uc3m | Universidad Carlos III de Madrid

Grado en Ingeniería en Tecnologías de Telecomunicación
(2017-2018)

Trabajo Fin de Grado

"Detección de factores de riesgo en la conducción mediante técnicas de análisis de datos"

Miguel González Vidal

Tutor

Julio Villena Román

Leganés, Septiembre 2018

Resumen

Este estudio tiene como objetivo analizar los efectos de diferentes distracciones bien diferenciadas en el rendimiento de los conductores. Para ello se utilizan técnicas de análisis de datos utilizando el programa Rapidminer, para analizar diferentes conjuntos de datos que contienen desde información biométrica de los usuarios, sobre la conducción realizada y las distracciones ocurridas (todo ello sincronizando el tiempo), hasta la valoración subjetiva de los voluntarios sobre la dificultad para la realización de la conducción ante dichas distracciones.

Durante el proyecto se toma especial consideración en mostrar las dificultades y los resultados de cada una de las etapas del tratamiento de los datos, hasta su posterior modelado y extracción de conclusiones. Dicho modelado se centra principalmente en técnicas de aprendizaje no supervisado.

Entre las conclusiones finales se ha podido determinar que:

-Los sujetos sometidos a distracciones, tienen cambios en sus medidas biométricas, más específicamente un aumento de las mismas, aunque no existe una diferencia entre las respuestas suficientemente diferenciada como para distinguir unas de otras a partir de los datos tomados. Asimismo la cantidad de veces que los sujetos a los que se les está aplicando una distracción (en la mayoría de los casos oral) pierden la vista de la carretera es significativamente mayor que la de los sujetos a los que no se les distrae.

También se ha podido encontrar que las distracciones percibidas como más frustrantes y que más esfuerzo requieren son, las distracciones cognitivas, presumiblemente porque obligan a emplear parte de la capacidad de análisis que se utilizaría en la conducción en resolver otros problemas, las distracciones sensorio-motoras, porque obligan a perder la vista de la carretera momentáneamente, obligando a adaptarse más rápidamente a cualquier cambio y la distracción resultante de la combinación de varias distracciones, presumiblemente por desbordar la capacidad de atención de los sujetos.

También se ha observado que cuando un sujeto cree que una prueba es difícil puntúa su rendimiento en dicha prueba de manera generosa, y cuando percibe una prueba como fácil puntúa su rendimiento de manera más baja. Esto es debido a la percepción de los sujetos de que realizan toda prueba correctamente y su percepción de sus propias capacidades es muy subjetiva.

Por último cabe destacar que no se ha percibido ninguna diferenciación entre los resultados de las pruebas aplicadas a los sujetos por cuestiones de género, siendo los grupos obtenidos por las técnicas de agrupamiento muy equilibrados en el número de hombres y mujeres.

Índice de Contenido

Índice de figuras

Índice de tablas

1. Introducción

A continuación se procederá a detallar la motivación que ha llevado a la elección de este trabajo de fin de grado, así como los objetivos que se pretenden conseguir con el mismo.

1.1 Motivación

Según la OMS (Organización Mundial de la Salud) los traumatismos causaron 5 millones de muertes en todo el mundo en el año 2015, un 27% de ellos fueron consecuencia de accidentes de tránsito. La mortalidad por estos accidentes a nivel mundial es de 18,3 defunciones por cada 100.000 habitantes, aunque es mucho más acusada en los países de ingresos bajos que en los de ingresos altos [1].

En España según las series históricas de accidentes con víctimas fallecidas, heridos graves y leves de la DGT (Dirección General de Tráfico), podemos observar que en nuestro país se producen entre 80.000 y 100.000 accidentes de tráfico anuales desde hace más de dos décadas [2].

1 - Accidentes con víctimas

Años	Total	Variación respecto al año anterior	Vías interurbanas	Variación respecto al año anterior	Vías urbanas	Variación respecto al año anterior
1993	79.925	-7.368	35.814	-3.307	44.111	-4.061
1994	78.474	-1.451	34.354	-1.460	44.120	9
1995	83.586	5.112	37.217	2.863	46.369	2.249
1996	85.588	2.002	37.434	217	48.154	1.785
1997	86.067	479	36.551	-883	49.516	1.362
1998	97.570	11.503	44.388	7.837	53.182	3.666
1999	97.811	241	44.784	396	53.027	-155
2000	101.729	3.918	44.720	-64	57.009	3.982
2001	100.393	-1.336	45.483	763	54.910	-2.099
2002	98.433	-1.960	44.871	-612	53.562	-1.348
2003	99.987	1.554	47.567	2.696	52.420	-1.142
2004	94.009	-5.978	43.787	-3.780	50.222	-2.198
2005	91.187	-2.822	42.624	-1.163	48.563	-1.659
2006	99.797	8.610	49.221	6.597	50.576	2.013
2007	100.508	711	49.820	599	50.688	112
2008	93.161	-7.347	43.831	-5.989	49.330	-1.358
2009	88.251	-4.910	40.789	-3.042	47.462	-1.868
2010	85.503	-2.748	39.174	-1.615	46.329	-1.133
2011	83.027	-2.476	35.878	-3.296	47.149	820
2012	83.115	88	35.425	-453	47.690	541
2013	89.519	6.404	37.297	1.872	52.222	4.532
2014	91.570	2.051	35.147	-2.150	56.423	4.201
2015	97.756	6.186	34.558	-589	63.198	6.775
2016	102.362	4.606	36.721	2.163	65.641	2.443

Tabla 1 Accidentes con víctimas, fallecidos 30 días, heridos graves y leves (Series históricas-DGT, 2016, últimos datos disponibles) [2]

La siniestralidad vial tiene como consecuencias además de las pérdidas de vidas humanas o de calidad de vida, que son lo más importante, una serie de costes asociados que lacran las economías. Dichos costes se pueden desglosar en tres amplias categorías según un estudio encuadrado en la acción COST 313, que realizó la Comisión Europea a principios de los noventa para revisar la forma en que los países Europeos estimaban los costes de los accidentes y realizar recomendaciones sobre cómo deberían cuantificarse [3]:

1. Los costes económicos directos: costes médicos, costes de reparación o reemplazo de vehículos dañados y costes administrativos.

2. Los costes indirectos: el valor de la capacidad productiva perdida como consecuencia de la muerte prematura, o de la incapacidad permanente o temporal causada por el siniestro.
3. El valor de la calidad de vida perdida: el valor de la pérdida de salud o pérdida de disfrute de la vida de la víctima, así como el dolor, aflicción y sufrimiento de la víctima y sus familiares.

La suma de las tres categorías de costes anteriores, proporciona el coste total por víctima en un accidente de tráfico o el valor total resultante de prevenir un accidente.

En España en el año 2011 la DGT en colaboración con la Universidad de Murcia, estimó los costes asociados a los accidentes de tráfico con víctimas utilizando el método de disposición de pago. Según dicha estimación, un fallecido supondría un coste de 1,4 millones de €, contabilizando los costes directos e indirectos, así como el valor de una vida estadística. De la misma manera se estimaron los costes asociados a un herido hospitalizado en 219.000€ y los de un herido no hospitalizado en 6100€. Dichas valoraciones actualizadas en 2016 tomando como referencia la variación nominal del producto interior bruto per cápita se recogen en la siguiente tabla.

Víctimas	Coste unitario (€ 2016)	Víctimas		Coste total € (2016)	
		Si solo se cuentan las contabilizadas por el sector transporte[1]	Si se cuentan las contabilizadas por los sectores transporte y salud[2]	Si solo se cuentan las contabilizadas por el sector transporte[1]	Si se cuentan las contabilizadas por los sectores transporte y salud[2]
Fallecidos	1.445.962	1.810	1.810	2.617.191.969	2.617.191.969
Heridos hospitalizados	226.190	9.755	20.542	2.147.672.481	4.646.391.586
Heridos no hospitalizados	6.300	130.635	477.022	787.281.090	3.005.364.917
				5.552.145.540	10.268.948.473

Tabla 2 Costes de los accidentes. (Principales cifras de siniestralidad, DGT, 2016) [3]

Como se puede observar los costes de los accidentes en 2016 se cifran en hasta 10.268 millones de euros, lo que supone aproximadamente el 1% del PIB de dicho año si se tienen en cuenta los datos de los sectores transporte y salud.

Dada la información anterior es comprensible que cada año se destinen millones de euros a las campañas de prevención (comunicación) y (en mucha menor medida) a la investigación.

Entre los proyectos de la DGT para 2018 se destinará una inversión de 1 millón de € en ayudas a la investigación, contando con 24 proyectos de investigación ya aprobados que se citan en el balance de siniestralidad de 2017 [4].

En el mismo documento también se aporta información sobre los principales factores contribuyentes en accidentes mortales o graves, los cuales se recogen en la siguiente tabla.

Balance de Seguridad Vial 2017

DGT Dirección General de Tráfico

Factores contribuyentes en Accidentes mortales o graves (1)

Factores	% de accidentes mortales y graves
Conducción distraída o desatenta	32%
Velocidad inadecuada	26%
Cansancio o sueño	12%
Alcohol	12%
Drogas	11%

Tabla 3 Factores contribuyentes en accidentes (Balance de seguridad vial, DGT 2017) [4]

Como se puede observar, el alcohol y las drogas causan un 12% y 11% de los accidentes en España respectivamente. El sueño y la fatiga tiene una influencia similar a los anteriores, causando un 12% de los accidentes. Mientras que una velocidad elevada y la conducción distraída se coronan como los principales factores causantes de accidentes con un 26% y un 32% de los accidentes mortales y graves.

1.2 Objetivos

El objeto de este trabajo será determinar las principales distracciones que influyen en el rendimiento de los conductores, aumentando así las posibilidades de un accidente. Para ello analizaremos mediante técnicas de minería de datos, algunos conjuntos de datos que contienen información sobre 68 voluntarios que conducen en un simulador sometidos a diferentes distracciones bien diferenciadas.

- En primer lugar se intentará observar si los sujetos producen una respuesta medible en sus medidas biométricas como consecuencia de las distracciones aplicadas y si existe una diferencia en las respuestas de éstos a las diferentes distracciones aplicadas.
- En segundo lugar se observará la percepción que estos voluntarios tienen de las distracciones aplicadas, para deducir cuáles de estas distracciones son consideradas como más difíciles de sobrellevar manteniendo la concentración en la tarea de conducción y por tanto son percibidas como más peligrosas, o en otras palabras, son factores de más peso en la causa de un accidente. Dicha tarea se realizará a partir de los análisis realizados a las encuestas de la escala NASA-TLX que contiene las puntuaciones subjetivas que cada sujeto ha otorgado a las distintas pruebas [17].

2. Estado del arte

En el siguiente capítulo se proporcionará una visión de las disciplinas de extracción de conocimiento a partir de volúmenes de datos, del origen y desarrollo de las mismas, de la normativa legal que afecta a la realización de un proyecto de minería de datos y una breve clasificación de las técnicas de minería de datos (posteriormente se explicarán en detalle las técnicas empleadas), así como una breve descripción de la elección del programa de minería de datos utilizado.

2.1 Extracción de conocimiento y Minería de Datos

La extracción de conocimiento de grandes volúmenes de datos está estrechamente relacionada con una disciplina conocida como KDD (Knowledge Discovery from Data), que se refiere al proceso no trivial de descubrir conocimiento e información útiles contenidos en un repositorio de información. Muchas veces se le llama también Minería de datos (Data Mining), aunque este término se refiere en realidad a una de las etapas del proceso de descubrimiento de conocimiento útil, que se refiere específicamente a aplicación de algoritmos y métodos de extracción de patrones en nuestros datos. También se puede llamar a esta disciplina, knowledge mining from data (minería de conocimiento a partir de datos), knowledge extraction (extracción de conocimiento), data/patterns analysis (análisis de datos/patrones), data archaeology (arqueología de datos), o data dredging (dragado de datos) [5].

El proceso del KDD puede descomponerse como una secuencia iterativa de pasos cuyas fases se detallan a continuación (aunque dichas fases varían ligeramente según la bibliografía empleada):

1. Limpieza de los datos: estos deben ser filtrados correctamente para eliminar el ruido, los datos inconsistentes, así como los datos ausentes, que pueden causar que los análisis proporcionen resultados incorrectos o inexactos.
2. Integración de los datos: consiste en combinar múltiples fuentes de datos para obtener una mayor cantidad de información.
3. Selección de los datos: Consiste en seleccionar únicamente los datos relevantes para la tarea de análisis que vamos a realizar.
4. Transformación y reducción de los datos: la transformación es el proceso por el cual convertimos los datos en formas más apropiadas para aplicar una determinada técnica de minería de datos. La reducción de los datos consiste en encontrar las características más significativas del conjunto dependiendo del objetivo buscado, se pueden utilizar por ejemplo técnicas para reducir el número de variables consideradas o emplear distintas representaciones de los datos.
5. Modelado: un proceso esencial en el cual se aplican algoritmos y métodos para extraer patrones de datos.
6. Evaluación y presentación de los patrones obtenidos: se trata de identificar los patrones que realmente representan conocimiento utilizando análisis estadísticos u otros métodos.

7. Interpretación de los resultados: análisis de los resultados obtenidos, y si este no es satisfactorio se repetirán aquellos pasos anteriores que se consideren oportunos.

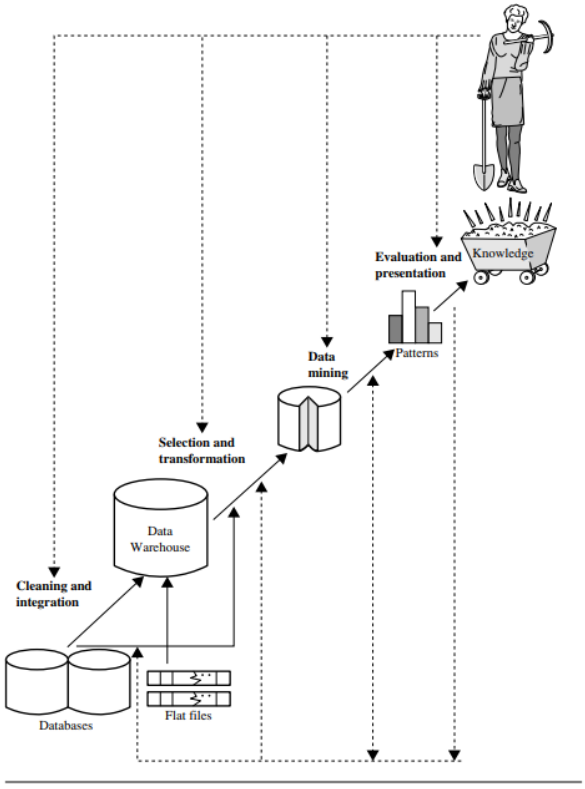


Fig. 1 Proceso del KDD (J. Han, M. Kamber & J. Pei, 2012) [5]

Además se pueden añadir dos pasos previos a los 7 primeros que se encuentran directamente relacionados con los anteriores:

1. Identificación de los objetivos: Definir cuáles son los objetivos buscados, para elegir un conjunto de datos adecuado a nuestro objetivo.
2. Selección de datos e información: Selección del conjunto o conjuntos de datos que contendrá las variables buscadas de cara a los objetivos seleccionados.

Actualmente sin embargo, la metodología más utilizada es el modelo CRISP-DM (Cross Industry Standard Process for Data Mining). Su modelo de referencia consta de 6 fases que se mostrarán a continuación [6]:

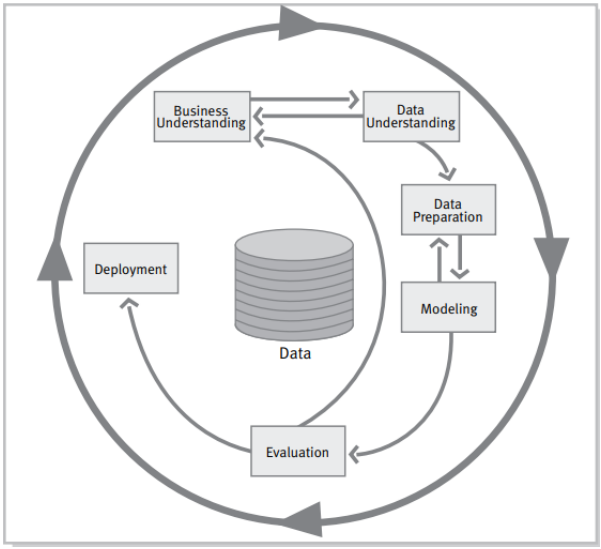


Fig. 2 Modelo CRISP-DM ("CRISP-DM 1.0") [6]

El siguiente diagrama muestra las 6 fases del proceso, las cuales no siguen siempre la misma secuencia, sino que el resultado de cada tarea es lo que determina la tarea a realizar posteriormente.

El círculo exterior simboliza que el proceso de minería de datos es una tarea cíclica que no finaliza con la implementación de la solución, sino que con esta y el conocimiento adquirido, surgen más preguntas a analizar.

A continuación se detalla el significado de las 6 fases:

- Entendimiento del negocio: Comprender los objetivos y requerimientos del negocio, para poder diseñar correctamente un plan con el que lograr dichos objetivos.
- Entendimiento de los datos: Fase en la que se busca obtener una familiaridad con los datos, para identificar problemas en la calidad de los mismos, obtener las primeras relaciones entre los datos y detectar segmentos interesantes a los que poder aplicar hipótesis.
- Preparación de los datos: En esta fase se busca obtener los conjuntos de datos finales que se utilizarán para realizar el modelado. Es un proceso que se realiza múltiples veces hasta obtener los resultados deseados.
- Modelado: Durante esta fase se seleccionan y aplican diferentes técnicas de modelado y se optimizan para obtener los resultados óptimos.
- Evaluación: En este punto se analiza si los modelos obtenidos, cumplen suficientemente con los objetivos para los que fueron creados, o si por otra parte alguno de los objetivos no ha sido suficientemente considerado.
- Despliegue: En este paso se busca utilizar la información obtenida de los modelos, o bien presentándola de manera que los clientes puedan entenderla o bien mediante la aplicación de modelos en vivo que permitan la toma de decisiones en tiempo real.

2.2 Origen de la metodología de extracción de conocimiento.

El origen del KDD ocurre de manera natural como una evolución de la tecnología de la información. Desde 1960 la tecnología de la información y de las bases de datos ha evolucionado desde primitivos sistemas de procesamiento a potentes sistemas de bases de datos [5].

En la década de 1970 la investigación y desarrollo progresó desde los primeros sistemas jerárquicos y bases de datos a bases de datos relacionales (donde los datos se almacenan en tablas de estructuras relacionales), herramientas de modelado de datos, indexado y métodos de acceso. Los usuarios ganaron un acceso mucho más flexible a los datos a través de lenguajes de consulta, interfaces de usuario y otras tecnologías.

Después del establecimiento de los sistemas de gestión de bases de datos, la tecnología de bases de datos comenzó a desarrollarse hacia los sistemas avanzados de bases de datos, almacenaje de datos, y técnica de minería de datos para análisis avanzado de bases de datos web.

A partir de mediados de la década de 1980 aparecieron los primeros sistemas de bases de datos avanzados. Dichos sistemas incorporaban nuevos y más potentes modelos de datos, como el modelo relacional extendido, orientado a objetos, bases de datos relacionales de objetos, y modelos deductivos. Además las bases de datos orientadas a aplicaciones favorecieron la creación de bases de flujo de datos de sensores, científicas y de ingeniería.

Las técnicas de análisis de datos avanzadas comenzaron a aparecer a finales de la década de 1980. Con el desarrollo del hardware de los ordenadores durante las tres últimas décadas permitieron grandes cantidades de ordenadores potentes y asequibles, equipo de toma de datos y medios de almacenamiento que proporcionaron un gran empujón a la industria de la información y bases de datos, permitiendo a un gran número de repositorios de información estar disponibles para gestión de transacciones, recuperación de información y análisis de datos.

Una arquitectura de repositorio de datos emergente es Data warehouse (almacén de datos), un repositorio de múltiples fuentes de información heterogénea organizado bajo un esquema unificado en un único lugar para facilitar la toma de decisiones.

Desde entonces grandes volúmenes de datos se comenzaron a acumular en todos estos sistemas. Durante la década de 1990, la red mundial y las bases de datos basadas en web, empezaron a aparecer. Las bases de datos globales basadas en internet, como WWW facilitaron la difusión de información.

Actualmente los sistemas de almacenamiento de datos han seguido evolucionando, vivimos en la "edad de la información". Terabytes y Petabytes de información se mueven diariamente en las redes de ordenadores, la red mundial y los sistemas de almacenamiento de los negocios, las redes sociales, la ciencia y la ingeniería y cualquier otro aspecto de la vida diaria. Por tanto es todo un desafío poder procesar y analizar los enormes volúmenes de datos de cara a obtener conclusiones que mejoren la toma de decisiones en cada uno de estos campos.

2.3 Breve clasificación de las técnicas de minería de datos

Antes de comenzar, es preciso apuntar que la minería de datos posee cientos de técnicas distintas que ayudan a modelar, para explicar o predecir el comportamiento de diferentes sucesos. En este trabajo solo se utilizarán una pequeña muestra de las mismas

(principalmente agrupamiento), pero antes de ello vamos a describir brevemente una clasificación de las mismas. Y posteriormente una descripción detallada del funcionamiento de las técnicas empleadas.

Las técnicas de minería de datos se pueden distinguir principalmente según el objetivo que se pretenda conseguir con la aplicación de las mismas, diferenciándose las técnicas predictivas, que están orientadas a estimar valores futuros de los datos, a partir de los datos conocidos, y las técnicas descriptivas, las cuales están orientadas a descubrir patrones en los datos y encontrar relaciones entre los atributos que forman el conjunto de datos. También existen técnicas auxiliares que están más enfocadas en la verificación [7].

También se puede establecer una segunda clasificación según el uso que se vaya a hacer de los datos, se distingue así entre aprendizaje supervisado y no supervisado. El aprendizaje supervisado trata de manera similar a las técnicas predictivas, de construir un modelo que permita expresar una variable particular (la variable etiquetada) en términos de las otras variables. El usuario puede elegir entre estimar, clasificar o predecir el valor de dicha variable como su objetivo. Las técnicas de aprendizaje no supervisado sin embargo, no tienen conocimiento a priori sobre la clase a estimar. El conjunto de datos es tratado para encontrar relaciones entre las variables de manera similar a las técnicas predictivas.

En el caso de este trabajo, se utilizará la clasificación de técnicas predictivas, descriptivas y auxiliares, siguiendo el esquema que se muestra a continuación, al que se le realizarán algunas adiciones.

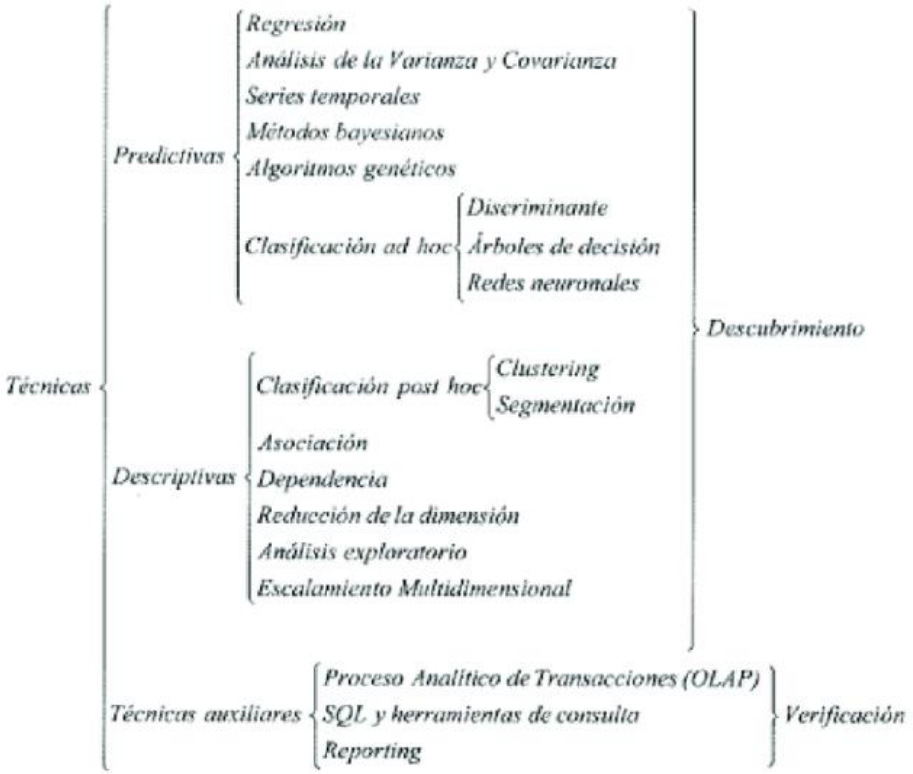


Fig. 3 Clasificación de las técnicas de minería de datos. C. Pérez y D. Santín. 2007 [7]

Técnicas predictivas: Como se ha expresado anteriormente los modelos predictivos intentan obtener los resultados futuros de una variable a partir del comportamiento de las demás. Dicha predicción puede ser un valor numérico o una categoría. Entre ellas se encuentran:

- **Clasificación:** Consiste principalmente, al igual que el aprendizaje supervisado, en la asignación de una categoría a nuevos datos sin clasificar, a partir del conocimiento obtenido de los datos anteriores. Típicamente se dispone de un conjunto de datos etiquetados y otro conjunto de datos nuevos sin clasificar. Cada muestra etiquetada se compone de un atributo etiqueta, que será el cual clasifique dicha muestra dentro de una de las posibles categorías, y de múltiples atributos de los que se obtendrá información para realizar las predicciones. El conjunto de datos nuevos dispondrá únicamente de los atributos que proporcionarán la información para la predicción y el objetivo es clasificar de la manera más precisa posible la pertenencia de cada muestra a su grupo adecuado. Algunos de los ejemplos más utilizados son: arboles de decisión, redes neuronales, y análisis de discriminante [8].
- **Regresión:** Esta técnica analiza la dependencia existente entre los valores de un atributo y los valores del resto de atributos de un conjunto de datos. El objetivo principal es construir un modelo para predecir los valores del primer atributo con el menor error posible. Se diferencia de la clasificación en que en la regresión el resultado obtenido es un valor continuo, mientras que en la clasificación es un valor discreto o una categoría. El modelo más utilizado es la regresión lineal, aunque también se utilizan modelos no lineales.
- **Análisis de varianza y covarianza:** Es un procedimiento estadístico que permite eliminar los efectos de oscurecimiento de las diferencias individuales preexistentes entre atributos, causados por la influencia de otra variable [9].
- **Análisis de series temporales:** consiste en la predicción de uno o más atributos con dependencia temporal, usualmente con el objetivo de obtener valores numéricos para predecir el comportamiento de sucesos con fuerte componente cíclica, como la predicción del tiempo, la bolsa, o predicciones demográficas. Para el análisis de dichos factores se utilizan métodos que ayudan a interpretar y visualizar los datos y extraer así información sobre las relaciones entre las distintas series de datos y poder predecir el comportamiento de la serie en momentos no observados [10].
- **Métodos Bayesianos:** modelo basado en el teorema de Bayes, que emplea las observaciones para actualizar la probabilidad de que una hipótesis sea o no cierta. Necesita información previa para determinar la distribución de probabilidad.
- **Algoritmos genéticos:** Intentan imitar cómo funciona la evolución biológica. Sometiendo a la población de individuos a acciones aleatorias (mutaciones genéticas) y después una selección de acuerdo a algún criterio, de manera que los más “aptos” sobreviven y los menos “aptos” son descartados [11].

Técnicas descriptivas: Están orientadas a describir el conjunto de datos que queremos analizar. Entre ellas se encuentran:

- **Técnicas de agrupamiento (clustering)** [21]**:** El objetivo de las técnicas de agrupamiento es clasificar los ejemplos de un set de datos formando grupos naturales (clusters) lo más homogéneos posibles entre los objetos que los componen y los más heterogéneos posibles con el resto de los grupos.

  Podemos distinguir dos grandes categorías de técnicas de agrupamiento según los métodos de agrupación:

  1. **Métodos Jerárquicos:** En cada paso del algoritmo solo un objeto cambia de grupo y los grupos están anidados en los de pasos anteriores. Si un objeto ha sido asignado a un grupo ya no cambia más de grupo. La clasificación resultante tiene un número creciente de clases anidadas.
  2. **Métodos No jerárquico o Repartición**: Comienzan con una solución inicial, un número de grupos **K** fijado de antemano y se agrupa los objetos para obtener los **K** grupos.

  Independientemente de la distinción mediante el método de agrupamiento, también hay que tener en cuenta la métrica de distancias o similitudes, que será el método por el que se calcule el parecido o la diferencia entre los objetos.

- **Recapitulación o resumen (Summarization):** Consiste en organizar la información en subconjuntos con descripciones simples para conseguir una abstracción o generalización de los datos. Para ello se utilizan medidas estadísticas como la media, la desviación típica, la varianza, la frecuencia, la moda y la media. Esto permite el visionado y estudio de los datos desde diferentes enfoques [8].

- **Reglas de asociación:** Consiste en descubrir hechos que ocurren en común dentro de un conjunto de datos [12]. Para encontrar reglas de asociación es necesario considerar todas las posibles combinaciones para que haya una consecuencia. De esta manera se establecen reglas que indican dependencias entre las muestras de un conjunto de datos.

- **Dependencias:** Se buscan dependencias probabilísticas o funcionales, causales o cuantitativas (fuerza de las dependencias). Se utilizan estos datos para dado el valor de un elemento predecir el valor de otro.

- **Reducción de la dimensión:** Se utiliza para reducir el número de atributos o variables a estudiar, es apropiado cuando se tienen muchos atributos con respecto al número de ejemplos del conjunto de datos.

- **Extracción de características:** La extracción de características está ligada a la reducción de la dimensión. Suele utilizarse cuando el conjunto de datos utilizado es muy extenso (y esto causa problemas con su procesamiento directo) y se sospecha que existe información redundante. Se procede entonces a transformar el conjunto de datos en un conjunto reducido, a lo cual se le llama selección de

características. El nuevo conjunto contiene variables que expresan la mayor parte de la información del antiguo conjunto con una representación reducida.

2.4 Aplicación de análisis de datos utilizada: Rapidminer

Actualmente existen múltiples herramientas de análisis de datos disponibles para los analistas, en este trabajo se expondrán las características principales de algunas de las más populares y posteriormente se detallará el software elegido para la realización del mismo, Rapidminer.

En la siguiente tabla comparativa se muestran algunas de las características principales de algunos de los programas cuyo uso está más extendido actualmente:

Programa	**Características principales**	**Lenguaje de programación**	**Licencia/precio**
Rapidminer	Alta capacidad de procesamiento Interfaz gráfica (facilidad de análisis sin programación) Gran cantidad de operadores de preprocesamiento y modelado Buenas herramientas de visualización de los datos Destaca en los análisis predictivos.	Java/C++	Versión Freeware reducida Diferentes versiones y precios
Weka	Completamente diseñado en java, muy portable. Gran cantidad de técnicas de preprocesamiento y modelado Fácil de usar por su interfaz gráfica de usuario No es muy potente en algoritmos de agrupamiento No destaca en la visualización de los datos	Java	Software Libre/gratuito
Orange	Destaca por su programación visual: se consigue una visualización de los datos rápida y sencilla que facilita la toma de decisiones. Aprende de las preferencias de los usuarios.	Núcleo del software C++ Lenguaje de entrada: Python	Software libre/gratuito
KNIME	Gran eficiencia en el tratamiento de los datos, extracción, transformación y carga. Segmentación en módulos: software orientado al flujo de datos	Java	Software libre/gratuito Permite la compra de paquetes de aplicaciones.

	Amplia gama de funciones: más de 1000 módulos y paquetes de aplicaciones preparados (ampliables mediante pago)		
SAS	Últimos avances en técnicas de predicción Visualización interactiva de datos. Escalabilidad Alto coste, solo adecuado para grandes empresas.	Lenguaje SAS	Software gratuito para instituciones públicas Licencia de alto coste, diferentes modelos de precios.

Tabla 4 Principales Software de análisis de datos (elaboración propia, 2018)

En el caso de este proyecto se ha elegido el programa Rapidminer debido a diferentes factores:

1. Su interfaz gráfica basada en operadores permite reducir la complejidad de la implementación de los algoritmos, permitiendo centrarse únicamente en el entendimiento del funcionamiento de los mismos y de los resultados que proporcionan.
2. Las herramientas de visualización que contiene permiten el entendimiento rápido del estado de los datos, de los resultados de los algoritmos y facilitan su exposición.
3. Cantidad de algoritmos: Rapidminer integra los programas de minería de datos weka, y R, así como algoritmos propios, obteniendo una gran cantidad de operadores disponibles para el usuario.
4. Permite importar datos desde múltiples fuentes, tablas de Excel, csv, archivos SPSS, bases de datos, así como trabajar en la nube.
5. Proporciona licencias de estudiante con todas las características desbloqueadas y capacidad de procesamiento de datos ilimitada.

Por todas estas características se considera Rapidminer el software más adecuado para la realización de este proyecto, posteriormente se mostrarán todos los pasos realizados en el programa para la implementación de los análisis, desde la carga y limpieza de los datos hasta la obtención de los resultados de los algoritmos.

2.5 Marco Regulador

En este apartado se proporcionará una breve visión de las normativas legales que hay que seguir cuando se realiza cualquier investigación en la que se realice tratamiento de datos en nuestro país.

En primer lugar los conjuntos de datos del estudio que se han utilizado, contienen múltiples informaciones de carácter personal, edad, sexo de los sujetos, e incluso videos completos de sus rostros mientras conducen (a través de los cuales se obtiene la posición de los ojos mediante reconocimiento de imágenes). El uso de dichos datos queda recogido por la ley Española de protección de datos de carácter personal que se fundamenta en el artículo 18 de la Constitución Española:

Constitución Española- Artículo 18.

"4. La ley limitará el uso de la informática para garantizar el honor y la intimidad personal y familiar de los ciudadanos y el pleno ejercicio de sus derechos."

Esta ley afecta a todos los datos que hacen referencia a personas físicas registradas sobre cualquier soporte, informático o no.

Según la ley Orgánica 15/1999, de 13 de diciembre, de Protección de Datos de Carácter Personal debemos cumplir las siguientes condiciones:

"Artículo 2. Ámbito de aplicación.

Se regirá por la presente Ley Orgánica todo tratamiento de datos de carácter personal:

a) Cuando el tratamiento sea efectuado en territorio español en el marco de las actividades de un establecimiento del responsable del tratamiento."

Dicha ley se aplica a este proyecto puesto que el tratamiento de dichos datos se está realizando en territorio español.

"Artículo 4. Calidad de los datos.

1. Los datos de carácter personal solo se podrán recoger para su tratamiento, así como someterlos a dicho tratamiento, cuando sean adecuados, pertinentes y no excesivos en relación con el ámbito y las finalidades determinadas, explícitas y legítimas para las que se hayan obtenido."

Los datos recogidos por el estudio están destinados a usarse en investigación de comportamientos de conducción bajo estímulos distractores cuidadosamente abstraídos, así como evaluación comparativa de canales fisiológicos y reconocimiento multiespectral de rostros. Por tanto este proyecto puede hacer uso de los datos dado que se adecúa a los objetivos para los cuales fueron recogidos

"Artículo 5. Derecho de información en la recogida de datos.

1. Los interesados a los que se soliciten datos personales deberán ser previamente informados de modo expreso, preciso e inequívoco:

a) De la existencia de un fichero o tratamiento de datos de carácter personal, de la finalidad de la recogida de éstos y de los destinatarios de la información."

Este artículo se cumple porque el propio estudio centrado en la recogida de los datos que usamos utilizó a 68 voluntarios que una vez informados de los datos que serían recogidos y de sus posteriores usos, dieron su consentimiento a la toma de los mismos y procedieron a la realización de las pruebas.

Una vez aclarada la aplicación de la ley de protección de datos queda por resolver el derecho al uso de los mismos, para lo cual se aplica el Real Decreto Legislativo 1/1996, de 12 de abril, por el que se aprueba el texto refundido de la Ley de Propiedad Intelectual. Según dicha ley debemos prestar atención a los siguientes artículos:

"Artículo 1. Hecho generador.

La propiedad intelectual de una obra literaria, artística o científica corresponde al autor por el solo hecho de su creación.

Artículo 2. Contenido.

La propiedad intelectual está integrada por derechos de carácter personal y patrimonial, que atribuyen al autor la plena disposición y el derecho exclusivo a la explotación de la obra, sin más limitaciones que las establecidas en la Ley."

"Artículo 12. Colecciones. Bases de datos.

1. También son objeto de propiedad intelectual, en los términos del Libro I de la presente Ley, las colecciones de obras ajenas, de datos o de otros elementos independientes como las antologías y las bases de datos que por la selección o disposición de sus contenidos constituyan creaciones intelectuales, sin perjuicio, en su caso, de los derechos que pudieran subsistir sobre dichos contenidos."

Según los dos primeros artículos, los datos del estudio constituyen una obra científica protegida por los derechos de autor, que atribuyen el derecho de uso y explotación únicamente a los autores de la misma. Además según el artículo 12, estos derechos son extensivos a las bases de datos que por su selección o disposición de contenidos constituyan una creación intelectual. Por tanto podremos usarlos solo con consentimiento de sus autores.

En nuestro caso los autores del conjunto de datos, los han puesto a disposición de cualquiera que quiera usarlos, a través de un proyecto público almacenado en la página web Open Science Framework, a continuación se muestran las capturas donde así se explica.

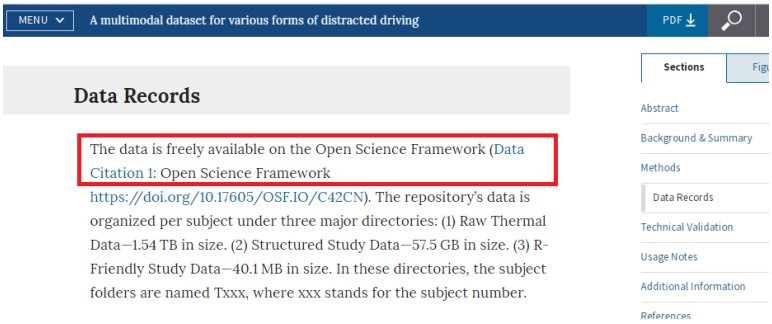


Fig. 4 Captura de permiso de uso de los datos en el repositorio (elaboración propia, 2018)

Fig. 5 Captura del permiso de uso de los datos de proyectos públicos en OSF (elaboración propia, 2018)

Por último solo queda decir que el uso del programa de análisis de datos utilizado, Rapidminer, está también regulado por la ley de propiedad intelectual, y se necesita una licencia apropiada para poder usarlo. Rapidminer ofrece licencias gratuitas con una capacidad limitada de procesamiento de datos (10.000 filas). Pero como nuestro proyecto necesitaba procesar casi 300.000 filas de datos inicialmente necesitaba adquirir otra licencia. Afortunadamente también ofrece licencias educacionales para los estudiantes, por lo que se obtuvo finalmente una licencia de 12 meses, con capacidad ilimitada de procesamiento de datos.

3. Obtención y descripción de los datos

En este capítulo se muestra la plataforma en la cual se han obtenido los datos del experimento, dado que la toma de datos es altamente costosa tanto en tiempo como en recursos económicos, y no se podía por tanto realizar una toma de datos específica para nuestro proyecto. También se describirá pormenorizadamente cada variable de los archivos utilizados y por último, se describirán las variables creadas a partir de los datos originales para obtener medidas de interés relativas a nuestros objetivos.

3.1 Obtención de los datos de estudio.

Los datos con los que se ha realizado el trabajo han sido obtenidos de una plataforma de creación de proyectos donde investigadores de todo el mundo pueden crear, gestionar y compartir datos y resultados con otros investigadores con intereses comunes.

Dicha página es Open Science Framework y el acceso concreto a los datos puede realizarse desde este link: https://osf.io/c42cn/files/ donde cualquiera puede descargarlos y utilizarlos al ser estos públicos [14].

A continuación se muestra una captura de los directorios de descarga de los datos (figura 6):

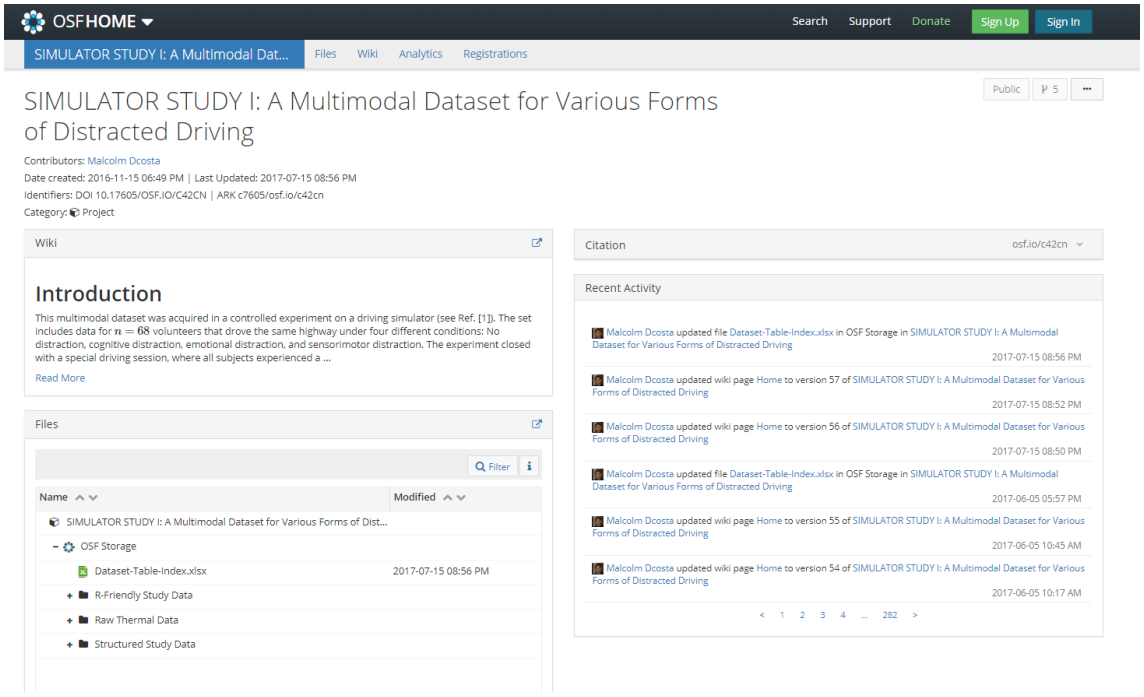


Fig. 6 Repositorio de datos utilizado (elaboración propia, 2016)

Los datos para estos análisis han sido obtenidos por el estudio de Malcolm Dcosta ""“SIMULATOR STUDY I: A Multimodal Dataset for Various Forms of Distracted Driving.”[14].

3.2 Descripción del experimento y de las variables utilizadas:

El experimento registra los datos de un simulador de conducción en el cual 68 voluntarios circulan por una misma autopista sometidos a distintas distracciones y en diferente orden cada vez, mientras esto ocurre se registran continuamente variables biométricas de los usuarios (ritmo cardíaco, respiración, posición de los ojos, etc…) así como medidas registradas por la propia máquina (aceleración, frenado, velocidad, distancia recorrida, etc…) [15].

A continuación se describirán con detalle todas las variables del experimento aunque algunas no se usarán en este proyecto.

En Primer lugar los 68 **archivos Txxx.b** (con xxx representando el número de tres cifras que identifica a cada sujeto), que contienen los datos biográficos siguientes:

- Género: Hombre/mujer
- Edad: Número correspondiente.
- Franja de Edad: (Joven o anciano).

En segundo lugar tenemos los 68 archivos **Txxx.bar** (con xxx siendo el número asociado a cada sujeto) que contienen la información psicométrica correspondiente a cada sujeto. Este archivo contiene las puntuaciones subjetivas que cada usuario ha puntuado a cada prueba, midiendo:

- Demanda mental percibida
- Demanda física percibida
- Demanda temporal
- Evaluación del rendimiento propio
- Evaluación del esfuerzo
- Frustración

Cada uno de estos atributos se evalúa con un número del 0 al 20, siendo 0 la menor puntuación y 20 la máxima.

A continuación se muestra una imagen con el test completo en el cuál se incluyen preguntas para aclarar cómo deben puntuar los sujetos cada uno de estos aspectos (fig. 9).

Dicha evaluación personal y subjetiva, se lleva a cabo mediante la escala NASA-TLX (National Aeronautics and Space Administration-Task Load Index), una herramienta que califica la carga de trabajo percibida para evaluar la efectividad de una tarea, sistema o equipo u otros aspectos del desempeño. Ha sido citada en más de 4400 estudios resaltando su influencia en la investigación de factores humanos [17].

Para realizar la medida, a los sujetos se les entregaba un test como el siguiente:

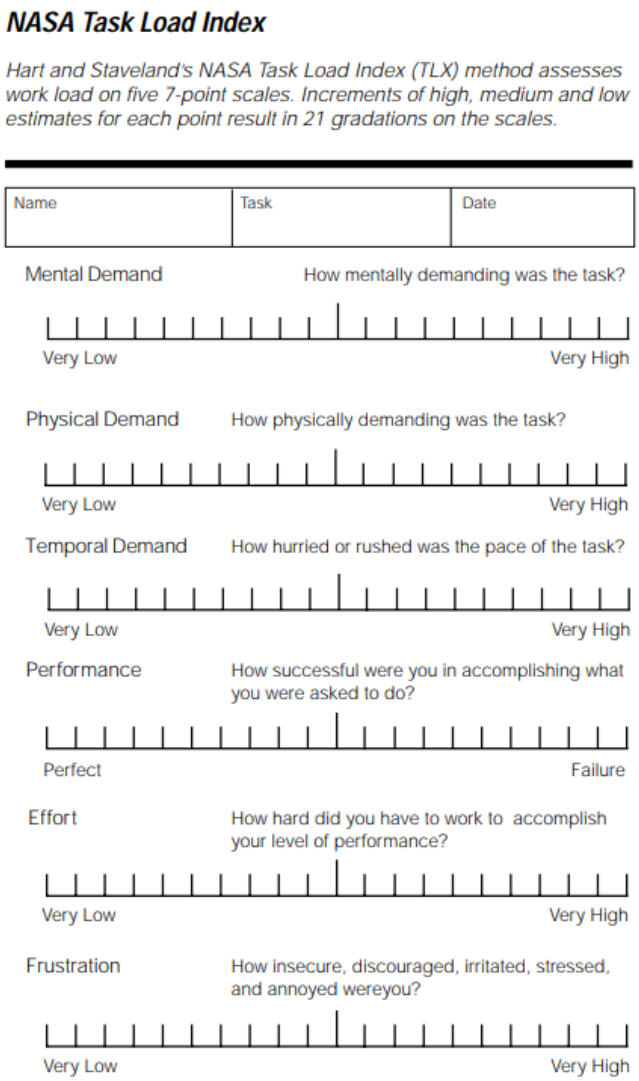


Fig. 7 Escala NASA-TLX, NASA Ames Research Center [17].

En el test NASA-TLX completo hay una segunda parte en la cual se comparan las medidas anteriores por pares, dando a elegir al sujeto cuál de ellas es más relevante para la realización de la tarea, el número de veces que una tarea es elegida es la puntuación de su peso, el cual es multiplicado por la puntuación de la escala por cada dimensión y dividido entre 15 para conseguir una puntuación de 0 a 100. De esta manera adquieren más importancia los datos que los usuarios han juzgado como más importantes, pero al ser esto una valoración subjetiva más, realizada sobre la anterior, muchos investigadores eliminan esta segunda parte del test, y usan solo la primera parte conocida como Raw-TLX, como es el caso de este estudio.

Por último tenemos los 68 archivos Txxx.csv (con xxx siendo el número asociado a cada sujeto) que contienen el registro de cada tipo de prueba de conducción, las fases durante las cuales se aplica un estímulo para distraer al sujeto, y así como múltiples medidas biométricas y los datos del simulador de conducción, todo ello siguiendo un marco de tiempo. A continuación se describirán de manera pormenorizada los campos de información que contiene el archivo.

- **Tiempo** (time): marco temporal por la cual se relacionan todos los otros datos registrados. Medido en segundos.

- **Conducción** (Drive): Diferencia la prueba de conducción que se está ejecutando en un momento concreto. El estudio estructura las sesiones en 8 pruebas distintas, Base de referencia (BL, baseline), conducción de práctica (PD, practice drive), conducción relajante (RD, Relaxing Drive), y conducción con fallo (FL, Failed Drive), la cual puede ser de dos tipos, conducción con fallo normal, FDN (Failed Drive Normal) o conducción con fallo cargada (FDL, Failed Drive Loaded). Dichas sesiones tienen un orden constante el cuál se muestra a continuación:

  BL⇨1, PD⇨2, RD⇨3, FDL o FDN⇨8.

  Las conducciones cargadas son conducción normal (ND, normal drive), Conducción Cognitiva (CD, Cognitive Drive), Conducción emocional (ED, Emotional Drive), Conducción sensomotora (MD, sensorimotor Drive) que tienen un orden aleatorio en el intervalo [4,7].

  A continuación se describirá cada una de ellas:

  El primer grupo de pruebas se realizan con orden constante y no llevan asociadas ningún estímulo externo para distraer al sujeto:

  - BL (Baseline): Punto de partida que sirve como referencia para las medidas de las otras pruebas de conducción. En este punto no se registran las medidas biométricas del sujeto. El sujeto mira a la pantalla del simulador, todavía vacía de información y se registra la posición de los ojos.
  - PD (Practice Drive): Fase diseñada para que el sujeto se familiarice con el simulador. Durante esta prueba se toman las medidas biométricas del sujeto, así como los datos del simulador.
  - RD (Relaxing Drive): Fase de conducción relajante, dedicada a tomar medidas en las cuales los sujetos se encuentren en estado de relajación.

    El segundo grupo de pruebas varía su orden para cada sujeto y además en estas pueden realizarse diferentes estímulos asociados con el carácter de la prueba para distraer al sujeto:

  - ND (Normal Drive): En esta fase se realiza una conducción normal sin estímulos añadidos.
  - CD (Cognitive Drive): Se fuerza al sujeto a pensar lógicamente, realizando preguntas analíticas o matemáticas de manera oral mientras este conduce. Mientras el sujeto es preguntado y hasta que responde se señaliza con un número en la variable estímulo, para conocer el intervalo de tiempo en el que el sujeto está dedicado a esta tarea.

- ED (Emotional Drive): Durante esta fase se realizan preguntas de manera oral dedicadas a sacar el lado emotivo de los sujetos. También queda señalizada la franja de tiempo durante la cual ocurren estas preguntas.
- MD (Sensorimotor Drive): Durante esta fase los sujetos intentan comunicarse mediante teléfono móvil, leyendo y contestando a mensajes y también llamadas.
- Por último se realiza una última fase en la que el simulador acelera solo, puede venir además acompañado de más estímulos externos, llamándose Conducción con Fallo normal (FDN) o conducción con fallo cargada (FDL), respectivamente.

- **Estímulo** (stimulus): Variable numérica que diferencia el estímulo o distracción que se está realizando en el momento actual:
  - 0: Sin estímulos.
  - 1: Preguntas analíticas.
  - 2: Preguntas matemáticas.
  - 3: Preguntas emotivas.
  - 4: Leer/escribir mensajes de texto.
  - 5: Mensajes de texto y llamadas.

- **Fallo mecánico** (Failure): Variable numérica que se utiliza para saber cuándo el vehículo está acelerando solo durante las pruebas FDN ó FDL.
  - 0: No hay aceleración por parte de la máquina.
  - 6: El vehículo del simulador acelera solo.

- **Actividad Electro-dérmica de las manos** (Palm.EDA): Se obtiene mediante un sensor que los sujetos llevan en las manos mientras realizan las diferentes pruebas de conducción. No se realiza toma de datos en la sesión de punto de partida. El valor de esta medida, al igual que las demás está sincronizado con el tiempo y su valor se mide en kΩ. Para ello se utiliza el sensor Shimmer3 GSR que puede medir un rango de 10 a 4.700 kΩ [15].

- **Transpiración Perinasal** (Perinasal.Perspiration): Obtenido mediante el procesamiento de las imágenes térmicas de la región perinasal. El algoritmo realiza un seguimiento de la región perinasal aunque el sujeto se mueva levemente, en el vídeo se reconocen diminutas "zonas frias" en las imágenes térmicas, las cuales corresponden a los poros abiertos del sujeto al sudar. Con esto se obtiene el valor estimado de la transpiración perinasal, la cual se mide en $^{0}C^{2}$.

- **Ritmo Cardíaco** (Heart.Rate): Para su medida se utiliza el sensor Zephyr BioHarnesss 3.0 (Zephyr Technology, Annapolis, MD) el cuál se adhiere a los sujetos con una correa debajo de la ropa y permite la toma del ritmo cardíaco y la respiración. La medida del ritmo cardíaco se registra en bpm y el aparato es capaz de detectar medidas entre 25 y 240 bpm [15].

- **Respiración** (Breathing.Rate): Como hemos mencionado en el apartado anterior, tanto el ritmo cardíaco como el número de respiraciones por minuto se toman con el mismo sensor, dicho aparato es capaz de registrar medidas en un rango de 4 a 70 bpm para el número de respiraciones por minuto.

- **Posición de los ojos** (Gaze.X.Pos y Gaze.Y.Pos): Se toma registro de la posición de los ojos mediante una cámara, cada una de las dos variables corresponde a una coordenada de la pantalla, horizontal y vertical respectivamente, y los casos en los que no hay registro de esta variable se considera una distracción visual, porque o bien el sujeto está mirando fuera de la pantalla, o bien tiene los ojos cerrados o está pestañeando.

- **Diámetro de las pupilas**: (Lft.Pupil.Diameter y Rt.Pupil.Diamenter): variable originalmente pensada para obtener mediante reconocimiento de imágenes el diámetro de la pupila de los sujetos y observar por tanto cuando se dilatan o se contraen las pupilas de los sujetos como un signo del estrés producido por las pruebas.

- **Distancia recorrida** (Distance): Es un atributo que nos permite ver la distancia recorrida por el coche en el simulador. Se mide en metros.

- **Velocidad** (Speed): Velocidad a la que circula el coche en el simulador en un instante dado. Se mide en Km/h.

- **Aceleración** (Acceleration): Aceleración del vehículo del simulador en el momento dado. El pedal está conectado a una válvula que registra el grado de utilización del mismo, de 0º (no se está pisando) a 90º grados (utilización total).

- **Fuerza en el freno** (Brake): Fuerza con la que el sujeto está pisando el pedal del freno del simulador. Se mide en Newtons.

- **Posición del volante** (Steering): Ángulo en el que se encuentra el volante en un momento dado, se mide en Radianes.

- **Posición respecto a la línea** (Lane.Position): Dicho atributo mide la posición del centro del coche respecto a la línea discontinua derecha de la autopista, que será la línea usada por los sujetos de este estudio. El carril mide 3,65m y el coche 1,85m. Los valores a la izquierda de esta línea son positivos y los valores a la derecha son negativos. Por tanto, los valores negativos significarían que el coche

está saliendo de la zona pavimentada de la carretera por la derecha, mientras que altos valores positivos, significarían que se está invadiendo el carril contrario. Altos valores positivos son valores superiores a 2,725m. Razón:(3.65m-1.85/2=2.725m).

- **Desplazamiento de carril** (LaneOffset): desplazamiento desde el centro del carril.

3.3 Variables añadidas por su utilidad para los análisis

Ahora describiremos las variables de apoyo que introduciremos para obtener medidas de interés:

- **EyesOut**: esta variable se utiliza para saber cuándo los sujetos apartan los ojos de la carretera. Tiene valor 1 cuando los sujetos apartan la vista de la pantalla y valor 0 cuando los sujetos están mirando la pantalla. Se basa en la idea de que la cámara de reconocimiento facial que hace seguimiento de los ojos está activa en todo momento y cuando esta no registra un valor de posición es cuando la cámara no reconoce los ojos porque el sujeto no está mirando a la pantalla o tiene los ojos cerrados (incluidos los pestañeos).

- **DistractionOn**: para saber cuándo se está aplicando una distracción sin diferenciar el tipo de estímulo que se produce. Es 1 cuando se está distrayendo al sujeto y 0 en caso contrario.

- **Abs.Inc.Heart.Rate**, **Abs.Inc.Breathing.Rate**, **Abs.Inc.Perinasal.Perspiration** y **Abs.Inc.Palm.EDA**: Miden el incremento o disminución absoluto respecto a la media de cada sujeto en estado de relajación, del ritmo cardíaco, la respiración, la transpiración perinasal y la actividad electro-dérmica de las manos respectivamente. Para ello se calcula la media de cada una de estas variables durante la prueba de conducción relajante. En cada magnitud se resta la media al valor actual para obtener el incremento o disminución absoluto de dicha variable respecto a su media.

- **Rel.Inc.Heart.Rate**, **Rel.Inc.Breathing.Rate**, **Rel.Inc.Perinasal.Perspiration** y **Rel.Inc.Palm.EDA**: Miden el incremento o disminución relativo respecto a la media de cada sujeto en estado de relajación, del ritmo cardíaco, la respiración, la transpiración perinasal y la actividad electro-dérmica de las manos respectivamente. Para ello se vuelve a utilizar la media de cada una de estas variables obtenida durante la prueba de conducción relajante. En cada magnitud se resta la media al valor actual y el resultado se divide por la misma media de cada sujeto para obtener el incremento o disminución relativos de dicha variable respecto a su media.

4. Preparación de los datos

Como ya se ha mencionado, en minería de datos el objetivo es obtener el conocimiento necesario acerca de alguna situación con el objetivo de controlarla o intentar predecir qué sucederá dadas ciertas condiciones conocidas. Usualmente dichos modelos se realizan a partir de cantidades enormes de datos, obtenidas de diferentes maneras, desde datos obtenidos por sensores automáticamente, hasta las respuestas de encuestas realizadas por personas y almacenadas en una base de datos.

Todos esos datos pueden contener múltiples errores, siguiendo con el ejemplo de los sensores, se pueden dar casos en los que se obtienen valores anormalmente altos (o bajos) por una falla en el propio sensor o por otra causa, dichos valores pueden modificar en gran medida el valor de la media de los datos, causando variaciones en los resultados de los algoritmos aplicados y creando por tanto, modelos que se alejan más del comportamiento real de las variables estudiadas (peores modelos). De la misma manera el sensor puede estropearse, desconectarse haber un obstáculo que bloquee parcial o totalmente las medidas, o que cambien repentinamente las condiciones en que funciona adecuadamente y dejar de tomar datos por una franja de tiempo en la que los demás sensores siguen registrando diferentes datos, y esa franja vacía debe ser rellenada. O en el caso del segundo ejemplo cuando se tratan datos obtenidos en una encuesta rellenada por personas, las palabras pueden estar escritas erróneamente, respuestas faltantes, o incompletas.

A lo anterior se añade además la tarea de añadir nuevos datos que son producto de los ya existentes, pero que se centren en un aspecto del cual queremos obtener información, como las variables mostradas en el capítulo 3.3.

Por tanto el analista pasará una gran parte del tiempo pre-procesando los datos para que los resultados obtenidos en los modelos tengan sentido. Esto es además un proceso reiterativo porque a medida que el analista vaya obteniendo resultados de los modelos empleados, se encargará de validarlos y ajustar aquellos aspectos que puedan desviar el resultado obtenido del resultado deseado.

A continuación se mostrarán algunas de las técnicas empleadas en el pre-procesamiento de datos y se realizarán en nuestro propio conjunto de datos, pero como se ha mencionado anteriormente esto es un proceso reiterativo, y se realizarán posteriores ajustes cuando sea necesario.

Como primer paso combinamos todos los documentos Txxx.csv en una única hoja de cálculo la cual llamaremos DatosFinales.csv. Podemos importarla en Rapidminer con el operador “Read Excel”, seleccionando la opción “Import Configuration Wizard…” seleccionamos la ruta donde se encuentra el archivo y permitimos que el programa tenga disponible la información de los datos para añadir más operadores posteriormente.

A partir de este momento además tenemos acceso a la pestaña de estadísticas de Rapidminer, la cual nos va a ayudar a identificar y a solucionar más fácilmente posibles errores en nuestros datos que reducirían la eficacia de los análisis que vamos a realizar.

A continuación vamos a realizar la primera preparación de los datos para deshacernos de los errores que podemos observar en pestaña de estadísticas de los datos, para ello vamos a centrar nuestra atención en tres criterios principalmente:

- Que los valores de los datos tengan sentido de acuerdo a la magnitud que miden: por ejemplo si obtuviésemos una magnitud negativa en el ritmo cardíaco de uno de los sujetos, la magnitud sería claramente errónea.
- Que los valores de los datos no superen el máximo o mínimo que pueden medir los sensores utilizados.
- Solucionar el problema de los datos que faltan: muchas técnicas de modelado no toleran los datos ausentes y es necesario rellenarlos o eliminar toda la fila de datos correspondiente para poder utilizarlas. Dependiendo de la cantidad de datos que falten en cada caso intentaremos utilizar la manera más óptima de solucionar el problema.

En la siguiente imagen (figura 8), se puede observar la hoja de estadísticas que nos proporciona Rapidminer al cargar nuestros datos. Se han marcado en azul los atributos a los cuales les faltan datos y en rojo los atributos cuyos valores no son coherentes con la magnitud medida o con los rangos de medida del sensor para mejorar su visibilidad.

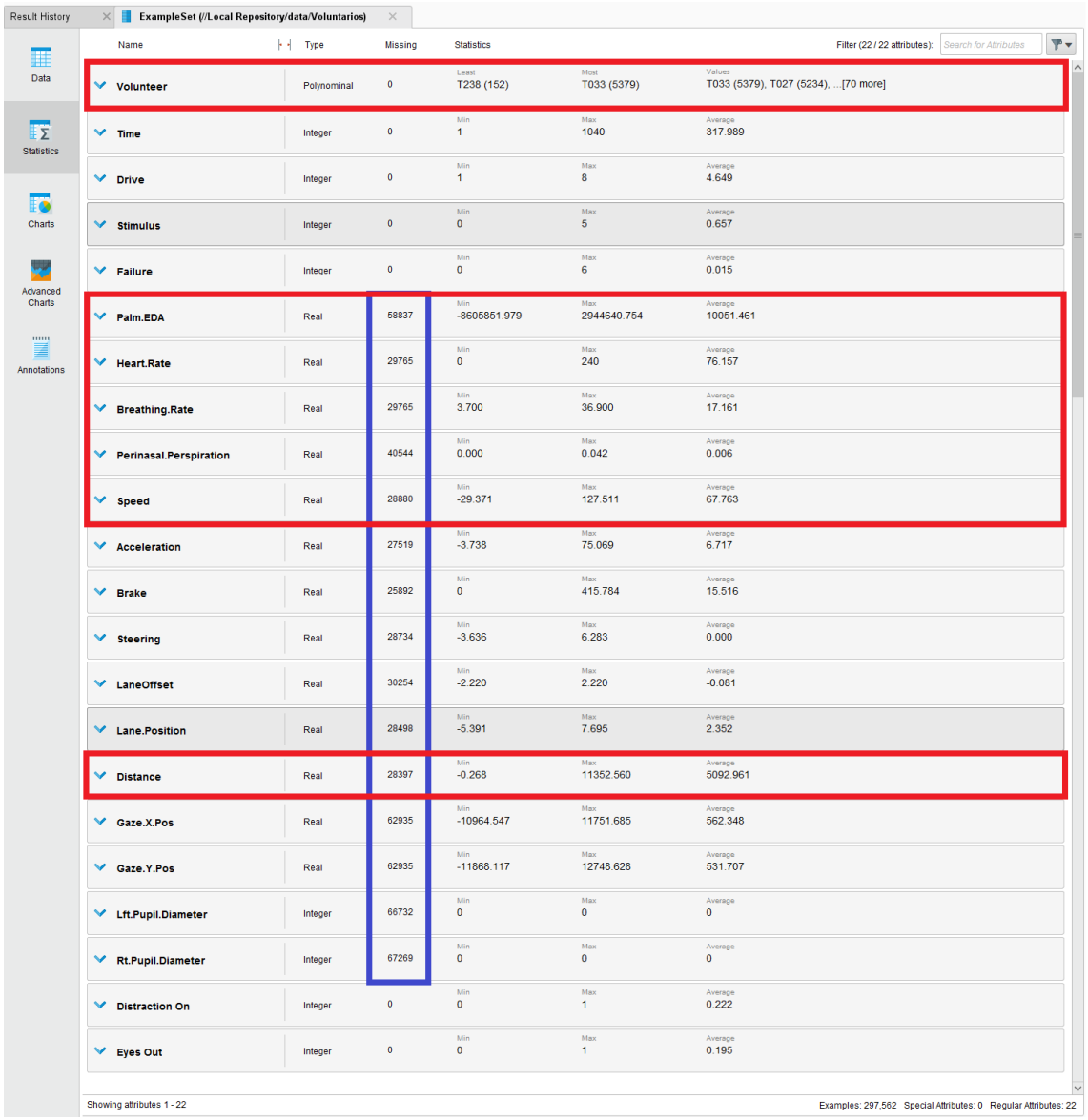


Fig. 8 Estadísticas de los datos antes de ser preprocesados (elaboración propia, 2018)

El primer error que podemos observar se encuentra en el atributo voluntario (volunteer), el estudio recogía los datos de 68 voluntarios y en nuestro programa aparecen 72 etiquetas distintas, por tanto en algún momento del proceso de recopilación de los datos en un único archivo hemos introducido un error por el cual aparecen 4 etiquetas más.

Si observamos la pestaña de valores de dicho atributo encontramos la siguiente información:

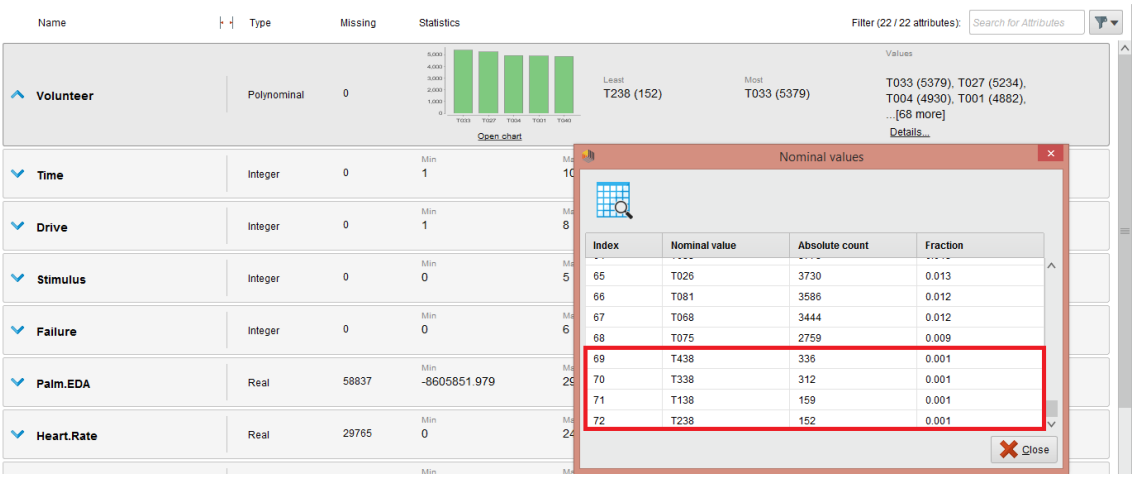


Fig. 9 Error en las etiquetas de los sujetos (elaboración propia, 2018)

Las etiquetas originales van en el rango de T001 a T088, con algunos saltos de números, pero todas tienen el formato T0XX, por tanto las etiquetas T438, T338, T238 y T138, son erróneas y como vemos su frecuencia de aparición es mucho menor a la de las etiquetas correctas. Por tanto se puede intuir que es un error al añadir la etiqueta de voluntario a cada archivo que originalmente se encontraba separado y juntarlos en uno solo.

Primero comprobamos que el archivo T038 efectivamente ha sido etiquetado erróneamente, y en segundo lugar comprobamos si el resto de datos del archivo han sido alterados de alguna manera. Tras comprobar que el resto de datos del archivo son los originales solo nos queda una cosa por hacer, arreglar las 4 etiquetas erróneas para que el programa los considere datos del sujeto T038 de nuevo, podemos hacerlo fácilmente en Excel, pero vamos a aprovechar este error para arreglarlo en el programa Rapidminer haciendo uso de las máximas prestaciones posibles del mismo.

Para ello comenzaremos creando un proceso en blanco, cargaremos el archivo de datos, y utilizaremos 4 operadores "Replace" consecutivos para reemplazar las etiquetas erróneas por la etiqueta T038, en las opciones seleccionamos single attribute, y el atributo voluntario para que únicamente reemplace en dicha columna, y por último escribimos las expresiones literales \bT138\b, \bT238\b, \bT338\b y \bT438\b en "qué reemplazar" y escribimos T038 en "por qué reemplazar". En la siguiente imagen se muestra el diagrama de bloques en Rapidminer.

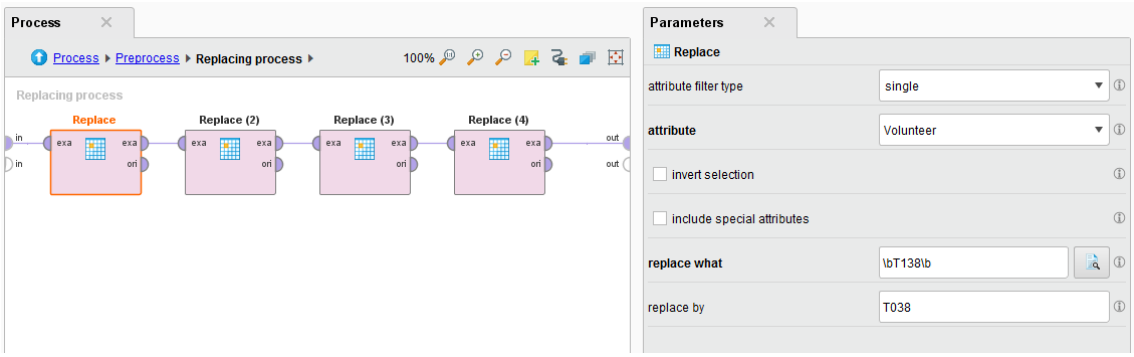


Fig. 10 Operadores replace y expresiones literales para corregir las etiquetas erróneas (elaboración propia, 2018)

Por último organizamos los cuatro operadores dentro de un subproceso al que llamaremos "Replacing process" y presionamos "start" para comenzar el proceso. Al observar de nuevo la pestaña de estadísticas vemos que el problema está solucionado.

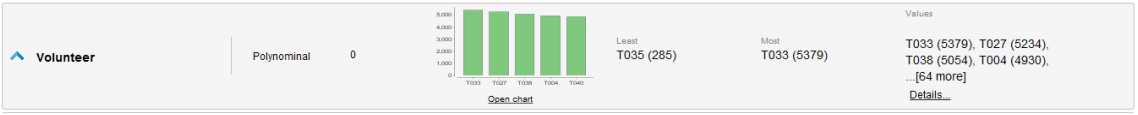


Fig. 11 Etiquetas de los sujetos corregidas (elaboración propia, 2018)

Volvemos a observar la tabla de estadísticas (figura 8). Las 4 siguientes variables tiempo, prueba de conducción, estímulo y fallo, parecen completamente correctas, no tienen valores ausentes, y los rangos de valores son consecuentes con la medida que indica la variable. Por lo que no requieren más atención.

Antes de continuar vamos a utilizar un filtro que elimine la primera prueba de conducción Baseline del conjunto de datos que vamos a utilizar (aunque guardaremos los datos no utilizados para posibles usos posteriores). En esta prueba el sujeto conduce por una autopista vacía en línea recta sin tener que hacer nada salvo mirar la pantalla, todavía no se toman las medidas biométricas, y por tanto faltan todos los datos de dichas columnas durante la prueba. Al eliminarla de este conjunto de datos podremos saber con más precisión cuantos datos faltan realmente (recordemos que los datos en la prueba baseline no se han tomado de manera intencional). Y conservaremos la información que buscamos para los análisis, porque dispondremos de la información de las pruebas "conducción de práctica", "conducción relajante" y "conducción normal" en las que el sujeto no es sometido a distracciones, y las cuatro pruebas donde el sujeto si es sometido a distracciones. A continuación se muestra el bloque añadido al proceso de Rapidminer:

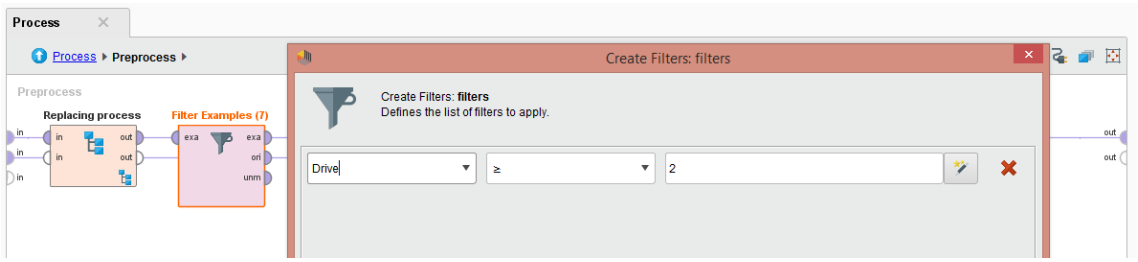
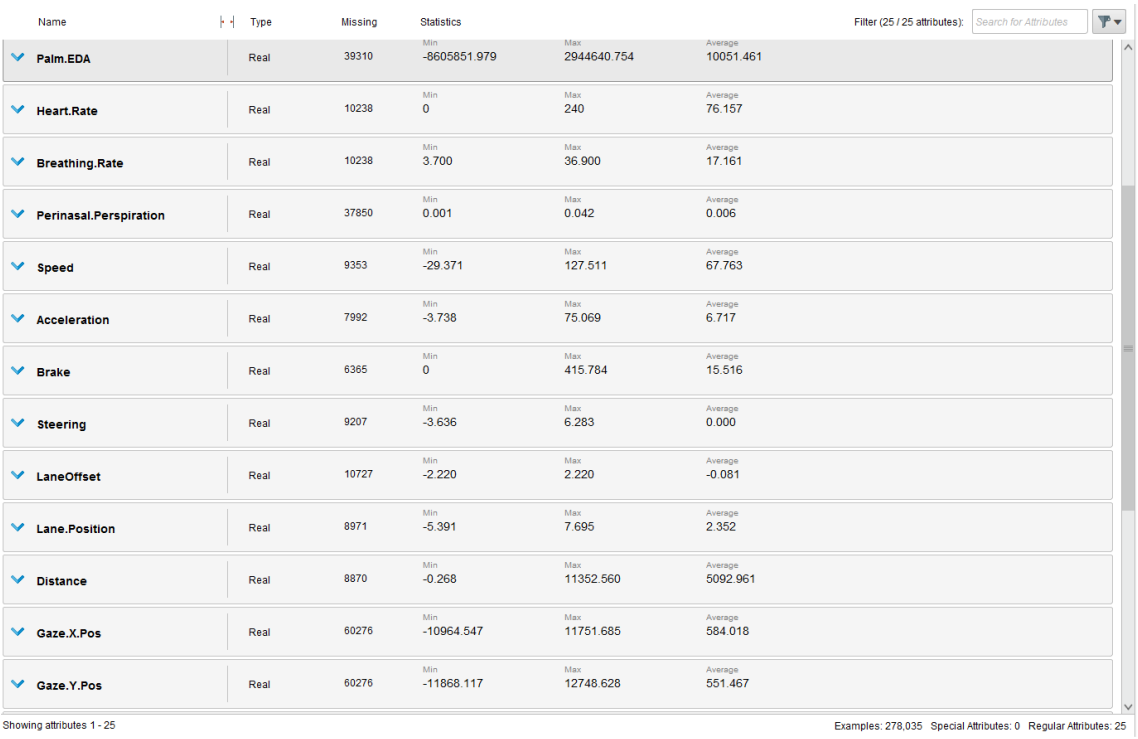


Fig. 12 Eliminación de los datos de la sesión BL (elaboración propia, 2018)

Y ahora observamos la reducción de valores ausentes casi todas las variables (figura 13).



Name	Type	Missing	Statistics		
			Min	Max	Average
Palm.EDA	Real	39310	-8605851.979	2944640.754	10051.461
Heart.Rate	Real	10238	0	240	76.157
Breathing.Rate	Real	10238	3.700	36.900	17.161
Perinasal.Perspiration	Real	37850	0.001	0.042	0.006
Speed	Real	9353	-29.371	127.511	67.763
Acceleration	Real	7992	-3.738	75.069	6.717
Brake	Real	6365	0	415.784	15.516
Steering	Real	9207	-3.636	6.283	0.000
LaneOffset	Real	10727	-2.220	2.220	-0.081
Lane.Position	Real	8971	-5.391	7.695	2.352
Distance	Real	8870	-0.268	11352.560	5092.961
Gaze.X.Pos	Real	60276	-10964.547	11751.685	584.018
Gaze.Y.Pos	Real	60276	-11868.117	12748.628	551.467

Filter (25 / 25 attributes): Search for Attributes

Showing attributes 1 - 25

Examples: 278,035 Special Attributes: 0 Regular Attributes: 25

Fig. 13 Reducción en los valores ausentes (elaboración propia, 2018)

Filtro De Valores Correctos

Al volver a observar la tabla de estadísticas, vemos que muchas de ellas tienen algunos valores sin sentido, o fuera del rango de valores que capta el sensor que las toma, de manera más pormenorizada observaremos:

- El sensor que mide la variable Palm.EDA devuelve valores que se encuentran entre 10 y 4.700 kΩ. Como se puede observar el valor mínimo es negativo, por lo que debe ser un error del sensor o del programa que registró los datos. Así mismo el límite superior del conjunto es muy superior al valor máximo que puede registrar el sensor.
- El sensor que mide ritmo cardíaco (Heart.Rate), es capaz de detectar entre 25 Bpm y 240bpm, el menor valor es inferior a dicho límite, aunque el valor máximo está dentro del límite superior.
- El sensor que mide la respiración (Breathing.Rate) detecta un ritmo respiratorio de entre 4 y 70 respiraciones por minuto: el mínimo en nuestro conjunto de datos se encuentra entre 3,7 y el máximo en casi 37 respiraciones por minuto.
- Por último la velocidad y la distancia tienen algún valor negativo, lo cual no tiene sentido, porque la distancia mide el número de metros recorridos y comienza en cero, y la velocidad solo tendría sentido si se considerase que el coche se mueve en dos sentidos (hacia delante y hacia atrás) pero todas las pruebas se realizan en una autopista y siempre se va hacia el frente. Además comprobando los pocos valores negativos de ambas (mediante un filtrado), estos solo corresponden al comienzo de algunas pruebas (tiempo igual a 1), donde en raras ocasiones el coche

comienza en distancia negativa (detrás de la línea de salida) y la velocidad es ligeramente negativa (el coche va hacia atrás).

Para solucionar estos problemas vamos a realizar un subproceso que llamaremos filtro de valores correctos cuya implementación se muestra a continuación:

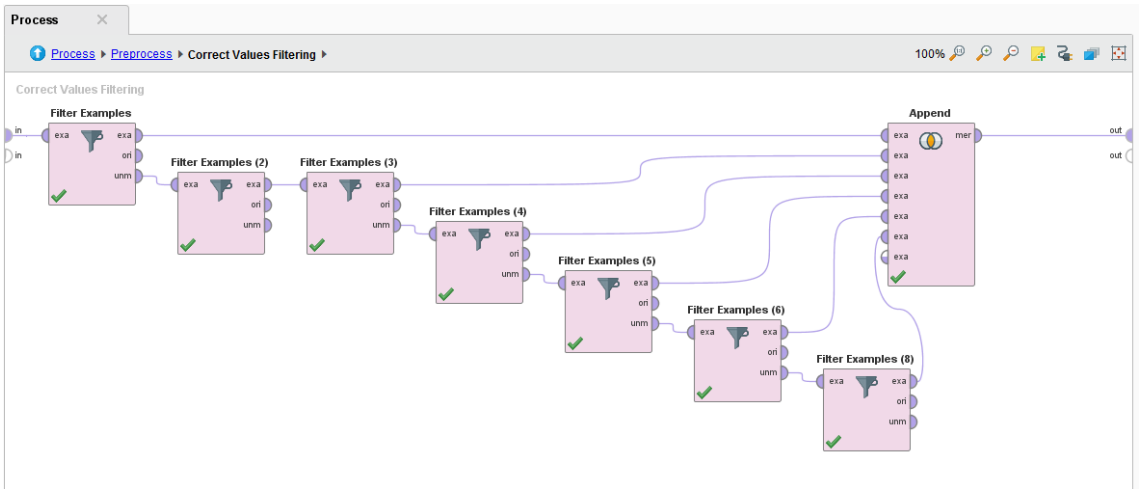


Fig. 14 Implementación de los filtros para eliminar valores incorrectos (elaboración propia, 2018)

Comenzaremos creando un primer filtro que separe todas las filas cuyos valores (todos a la vez) se encuentren dentro del rango de detección de los sensores.

Utilizamos el operador "filter examples" y la rellenamos como se muestra a continuación:

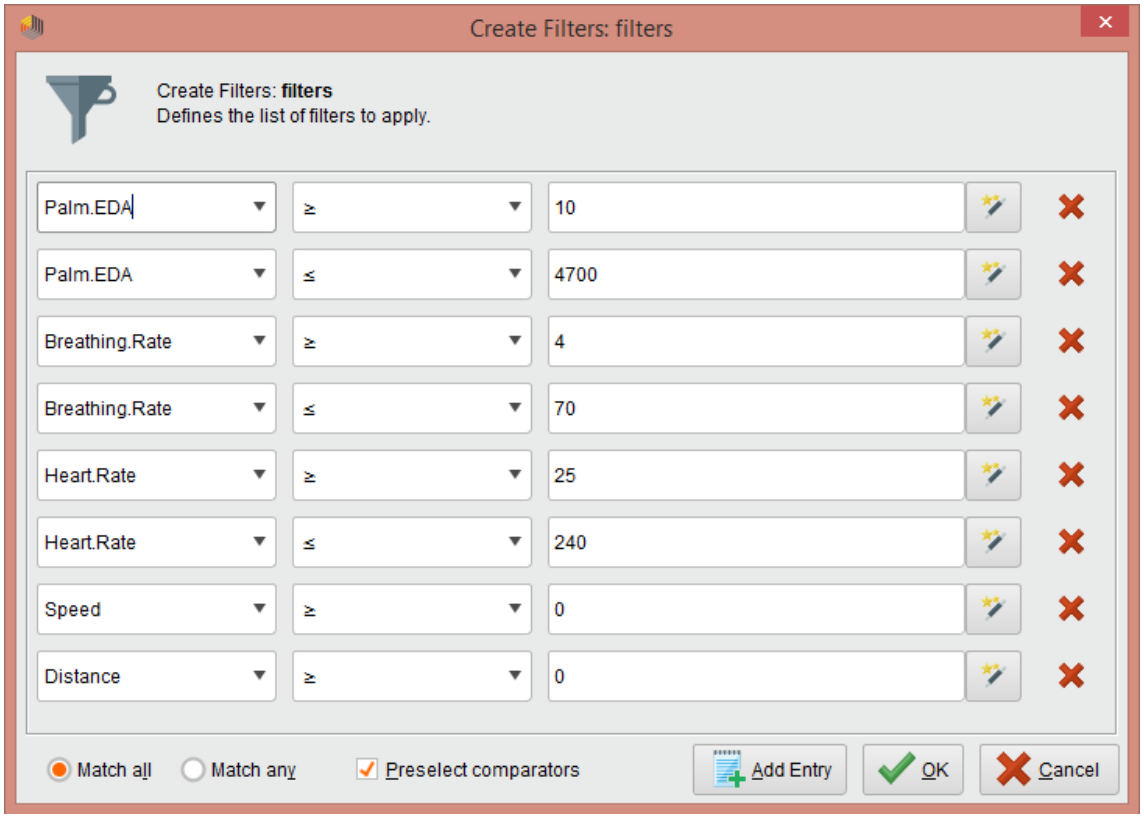


Fig. 15 Primer filtro (elaboración propia, 2018)

Si lo dejásemos como está ahora, no solo filtraría los valores incorrectos, sino que también perderíamos toda la fila de datos correspondiente a cualquiera de los atributos filtrados si uno de estos tiene un atributo ausente. Eso nos dejaría con aproximadamente 160.000 ejemplos de los 278.000 originales, aproximadamente perderíamos un 43% de la información.

Para que esto no ocurra vamos a conservar los datos que han sido eliminados porque tenían información ausente. Conectamos la salida "unmatched" del primer filtro a un nuevo filtro que recogerá las demás variables cuando alguno de los atributos que hemos filtrado tiene valor ausente. Añadimos una entrada para Palm.EDA, Breathing.Rate, Heart.Rate, Speed y Distance y seleccionamos "is missing". Pulsamos "Match any" y le damos a OK. Ahora hemos recuperado los ejemplos descartados por tener valores ausentes. Pero no podemos reintroducirlos en el set directamente por dos motivos, el primero es que algunas de las filas que han sido descartadas por tener valores ausentes, pueden contener además un valor fuera de rango de detección de los sensores. Y el segundo caso que no queremos recuperar son las filas que tienen no solo uno sino varios atributos con valores ausentes, entorpeciendo más por introducir mucha falta de información, que la información que nos van a aportar los atributos restantes.

Para solucionar esto, vamos a introducir 5 filtros en cascada, que van a filtrar las filas eliminadas por tener atributos con valores ausentes que hemos recuperado con el filtro número dos. A continuación se muestra el primero de ellos (filter examples 3):

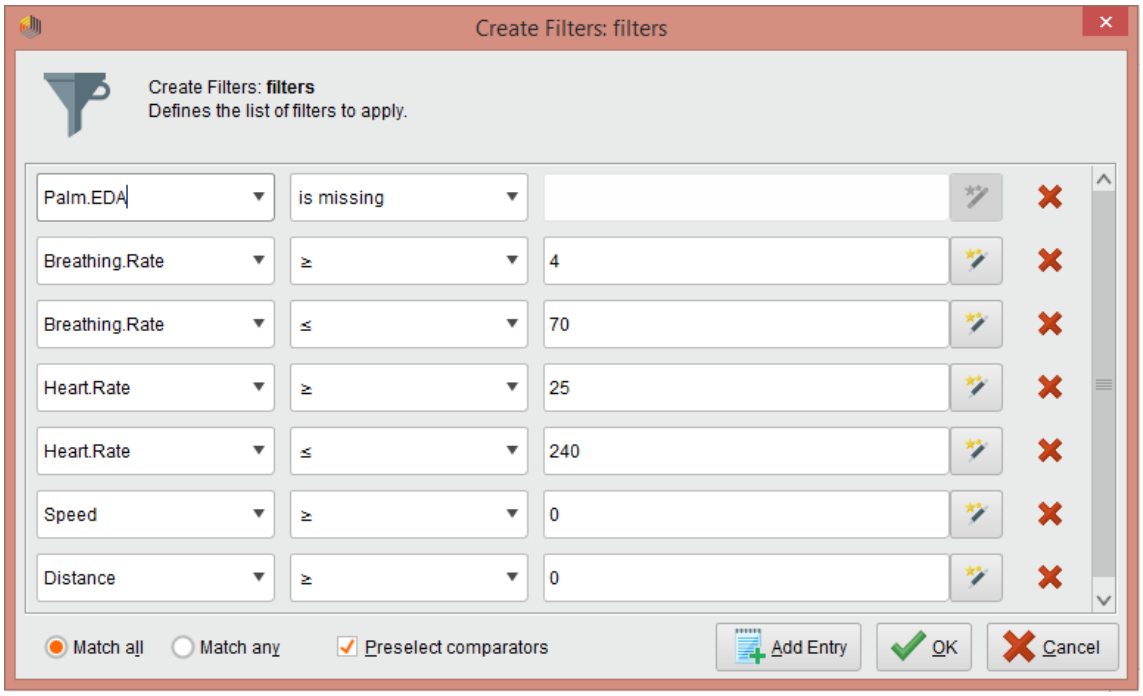


Fig. 16 Filtro para la recuperación de datos con solo un atributo ausente (elaboración propia, 2018)

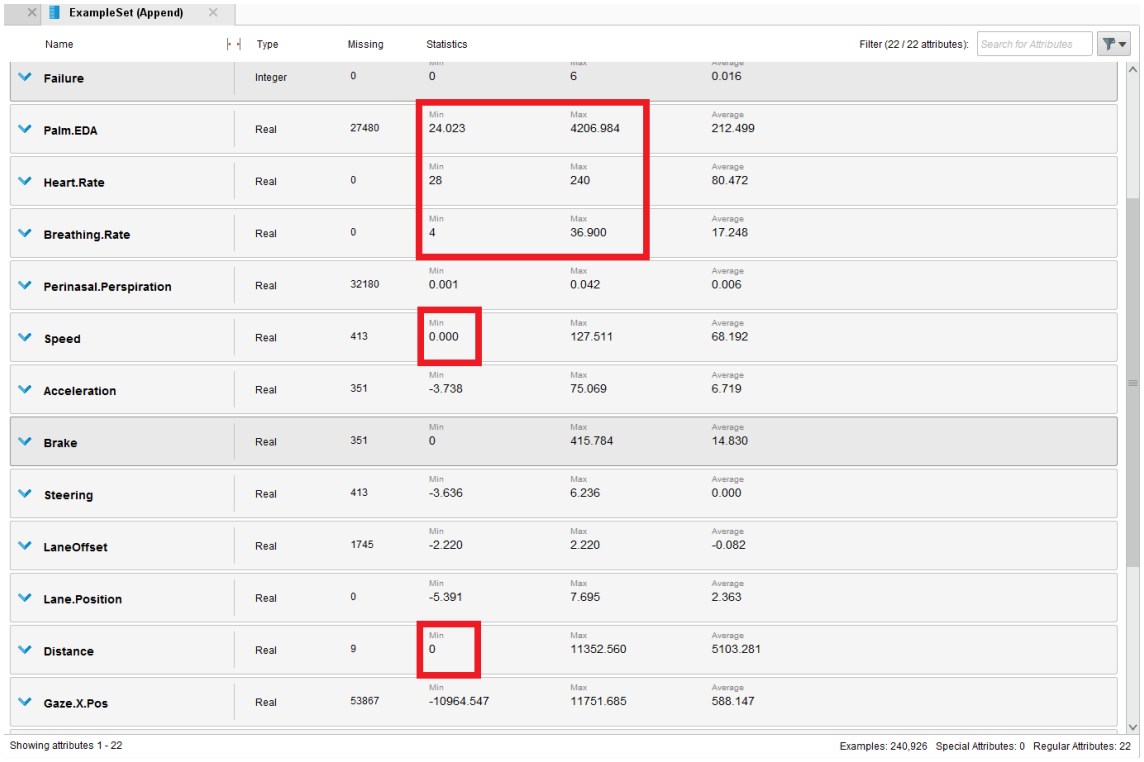
Las filas que no coincidan con esta medida, serán recogidas en la salida "unmatched", que se conectará al siguiente filtro.

En el siguiente filtro, Breathing.Rate tendrá valor ausente y los demás atributos se encontraran en el rango de medida de los sensores. Los atributos que no coincidan

volverán a ser filtrados en el siguiente filtro, que contendrá Heart.Rate ausente y los demás atributos en el rango de medida de los sensores.

De esta manera tras realizar los 5 filtros (figura 16), recuperaremos las filas que no tengan más de un atributo (de las medidas biométricas) ausente, eliminando así los grupos de datos sin ninguna medida biométrica que se pueden observar mirando la salida de datos sin filtrar. Y además entre estos no recuperaremos ninguno de los datos que hemos filtrado por encontrarse fuera de los valores de medida de los sensores.

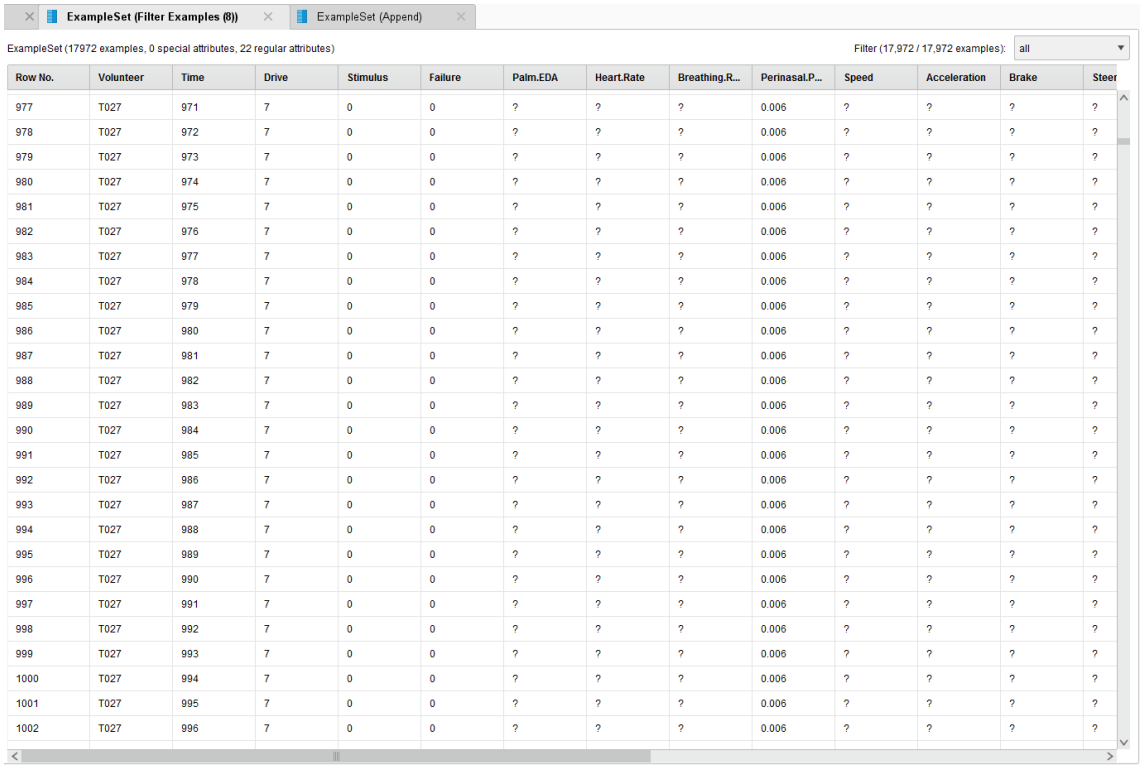
Al ejecutar el programa observamos que los atributos han corregido su rango de valores aunque todavía quedan muchos atributos con valores ausentes. Además podemos ver que conservamos aproximadamente 241.000 de los 278.000 datos originales, hemos perdido aproximadamente un 13% de las filas de datos.



Name	Type	Missing	Statistics		
Failure	Integer	0	0	6	0.016
Palm.EDA	Real	27480	Min 24.023	Max 4206.984	Average 212.499
Heart.Rate	Real	0	Min 28	Max 240	Average 80.472
Breathing.Rate	Real	0	Min 4	Max 36.900	Average 17.248
Perinasal.Perspiration	Real	32180	Min 0.001	Max 0.042	Average 0.006
Speed	Real	413	Min 0.000	Max 127.511	Average 68.192
Acceleration	Real	351	Min -3.738	Max 75.069	Average 6.719
Brake	Real	351	Min 0	Max 415.784	Average 14.830
Steering	Real	413	Min -3.636	Max 6.236	Average 0.000
LaneOffset	Real	1745	Min -2.220	Max 2.220	Average -0.082
Lane.Position	Real	0	Min -5.391	Max 7.695	Average 2.363
Distance	Real	9	Min 0	Max 11352.560	Average 5103.281
Gaze.X.Pos	Real	53867	Min -10964.547	Max 11751.685	Average 588.147

Fig. 17 Estadísticas de los datos conservados tras el filtrado (elaboración propia, 2018)

Mirando los datos que no mantendremos en el conjunto vemos que la mayoría de los que hemos eliminado corresponden a grupos de datos con enormes faltas de información. O bien todas las medidas biométricas, o bien todos los datos del simulador. Y en algunos casos ambos. Por tanto hemos eliminado muchas filas de datos con poca o nula información. Como se muestra en la siguiente captura:



ExampleSet (Filter Examples (8)) | ExampleSet (Append)

ExampleSet (17972 examples, 0 special attributes, 22 regular attributes) — Filter (17,972 / 17,972 examples): all

Row No.	Volunteer	Time	Drive	Stimulus	Failure	Palm.EDA	Heart.Rate	Breathing.R...	Perinasal.P...	Speed	Acceleration	Brake	Steer
977	T027	971	7	0	0	?	?	?	0.006	?	?	?	?
978	T027	972	7	0	0	?	?	?	0.006	?	?	?	?
979	T027	973	7	0	0	?	?	?	0.006	?	?	?	?
980	T027	974	7	0	0	?	?	?	0.006	?	?	?	?
981	T027	975	7	0	0	?	?	?	0.006	?	?	?	?
982	T027	976	7	0	0	?	?	?	0.006	?	?	?	?
983	T027	977	7	0	0	?	?	?	0.006	?	?	?	?
984	T027	978	7	0	0	?	?	?	0.006	?	?	?	?
985	T027	979	7	0	0	?	?	?	0.006	?	?	?	?
986	T027	980	7	0	0	?	?	?	0.006	?	?	?	?
987	T027	981	7	0	0	?	?	?	0.006	?	?	?	?
988	T027	982	7	0	0	?	?	?	0.006	?	?	?	?
989	T027	983	7	0	0	?	?	?	0.006	?	?	?	?
990	T027	984	7	0	0	?	?	?	0.006	?	?	?	?
991	T027	985	7	0	0	?	?	?	0.006	?	?	?	?
992	T027	986	7	0	0	?	?	?	0.006	?	?	?	?
993	T027	987	7	0	0	?	?	?	0.006	?	?	?	?
994	T027	988	7	0	0	?	?	?	0.006	?	?	?	?
995	T027	989	7	0	0	?	?	?	0.006	?	?	?	?
996	T027	990	7	0	0	?	?	?	0.006	?	?	?	?
997	T027	991	7	0	0	?	?	?	0.006	?	?	?	?
998	T027	992	7	0	0	?	?	?	0.006	?	?	?	?
999	T027	993	7	0	0	?	?	?	0.006	?	?	?	?
1000	T027	994	7	0	0	?	?	?	0.006	?	?	?	?
1001	T027	995	7	0	0	?	?	?	0.006	?	?	?	?
1002	T027	996	7	0	0	?	?	?	0.006	?	?	?	?

Fig. 18 Datos eliminados mediante el filtrado con múltiples campos ausentes (elaboración propia, 2018)

Ahora además añadiremos un filtro para eliminar duplicados por si hubiese datos dobles en el set de datos (y como comprobación por si hubiésemos añadido datos duplicados con nuestro filtro) y vemos que no hay filas duplicadas en el set de datos.

Antes de continuar vamos a introducir un filtro para eliminar algunos atributos de los cuales no vamos a obtener información.

En primer lugar eliminamos los dos atributos correspondientes a la dilatación de las pupilas Lft.Pupil.Diameter y Rt.Pupil.Diameter los cuales no contienen información. Están vacíos porque fueron creados para ser obtenidos mediante reconocimiento de imágenes a través de los videos de la cara de los sujetos durante las simulaciones, pero como eso no es objeto de este trabajo eliminamos las variables vacías.

En segundo lugar eliminamos las variables correspondientes a las coordenadas X e Y de la pantalla donde apuntan los ojos de los sujetos en todo momento. En vez de utilizar esa variable, utilizaremos la variable EyesOut, que muestra cuando se pierde visión de la carretera.

Ahora procederemos a ocuparnos de los múltiples valores ausentes del set de datos.

Tratamiento de los datos ausentes.

Si vamos mirando uno a uno los datos correspondientes a cada sujeto tras el filtrado, vemos que en la mayoría de ellos tienen un pequeño porcentaje de datos ausentes en un

atributo, salvo en algunos pocos casos donde todos los datos de un atributo faltan, más concretamente esto ocurre con los atributos Palm.EDA y Perinasal.Perspiration.

Una técnica aceptable para rellenar los huecos de valores ausentes cuando estos no son un gran porcentaje de los ejemplos totales es sustituir los mismos por la media de los datos.

En nuestro caso vamos a dividir los datos totales, separando los datos de cada sujeto para sustituir los huecos con valores ausentes por la media del atributo en cuestión de ese sujeto, y no por la media de los datos globales, que sería más imprecisa. Esto además no corregirá los datos de los sujetos con atributos que no tienen ninguna información, dándonos posteriormente la oportunidad de ocuparnos de ellos de la manera más apropiada posible.

Para realizarlo vamos a construir un nuevo subproceso, al que llamaremos "Missing value handling" que separará cada uno de los sujetos de los que se tiene datos y sustituirá los valores ausentes de cada atributo por la media de ese atributo para ese sujeto. Una vez hecho eso para todos los sujetos, se volverán a unir los datos en una sola tabla mediante el operador Append.

A continuación se muestran el subproceso.

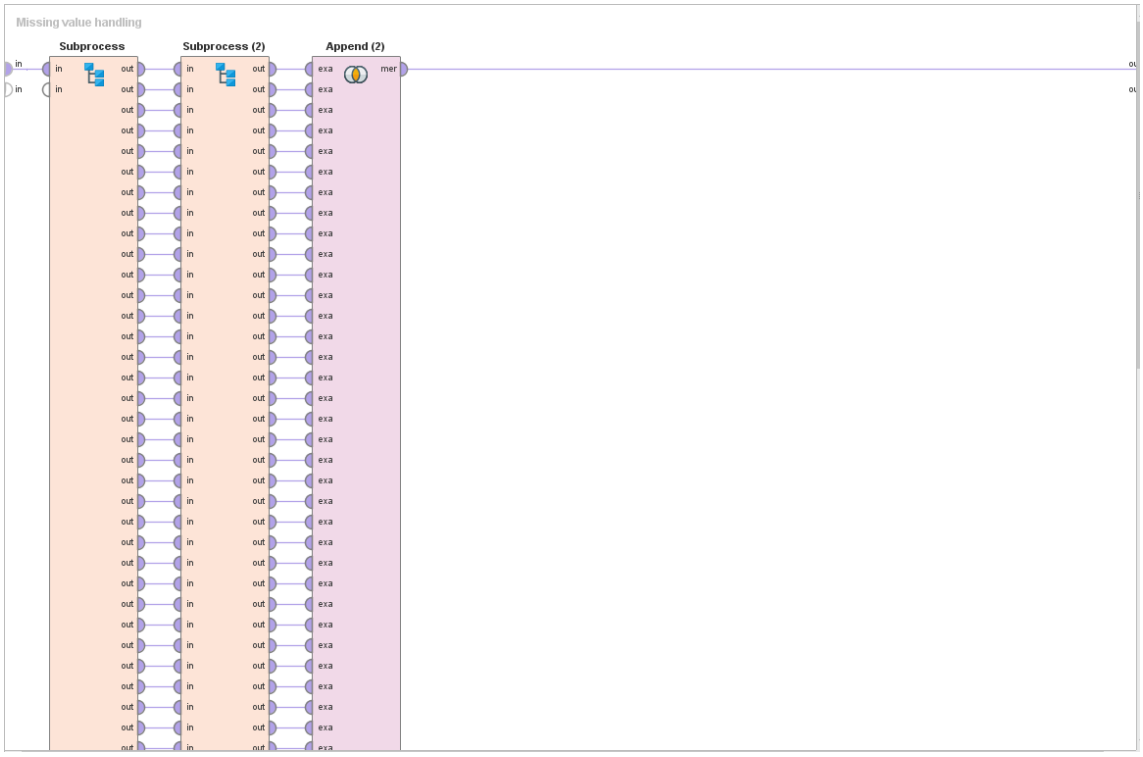
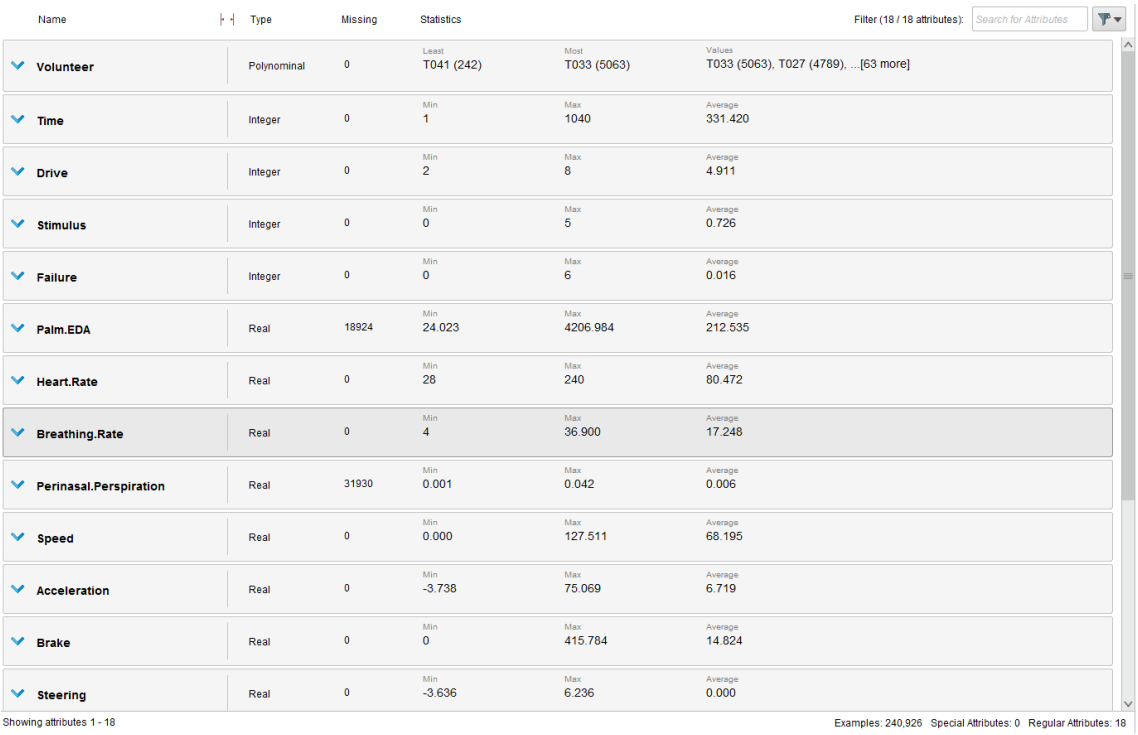


Fig. 19 Implementación del tratamiento de los valores ausentes (elaboración propia, 2018)

Internamente el primer subproceso tiene 68 operadores "filter examples", que envían a salidas diferentes los datos de cada sujeto (68 en total). En el subproceso cada entrada se conecta con un operador "Replace Missing values" y los datos rellenados se envían por salidas diferentes. Estas se vuelven a unir con un operador Append al final del subproceso.

Al ejecutar el proceso vemos que se han corregido todos los valores ausentes a excepción de 19.000 datos de la actividad electro-dérmica de las manos y 32.000 datos de transpiración perinasal.



Name	Type	Missing	Statistics		
Volunteer	Polynominal	0	Least: T041 (242)	Most: T033 (5063)	Values: T033 (5063), T027 (4789), ...[63 more]
Time	Integer	0	Min: 1	Max: 1040	Average: 331.420
Drive	Integer	0	Min: 2	Max: 8	Average: 4.911
Stimulus	Integer	0	Min: 0	Max: 5	Average: 0.726
Failure	Integer	0	Min: 0	Max: 6	Average: 0.016
Palm.EDA	Real	18924	Min: 24.023	Max: 4206.984	Average: 212.535
Heart.Rate	Real	0	Min: 28	Max: 240	Average: 80.472
Breathing.Rate	Real	0	Min: 4	Max: 36.900	Average: 17.248
Perinasal.Perspiration	Real	31930	Min: 0.001	Max: 0.042	Average: 0.006
Speed	Real	0	Min: 0.000	Max: 127.511	Average: 68.195
Acceleration	Real	0	Min: -3.738	Max: 75.069	Average: 6.719
Brake	Real	0	Min: 0	Max: 415.784	Average: 14.824
Steering	Real	0	Min: -3.636	Max: 6.236	Average: 0.000

Filter (18 / 18 attributes): Search for Attributes

Showing attributes 1 - 18

Examples: 240,926 Special Attributes: 0 Regular Attributes: 18

Fig. 20 Últimos datos ausentes (elaboración propia, 2018)

Al separar y observar solo el grupo de datos con valores ausentes, se puede observar que dichos datos pertenecen a sujetos a los que le falta solo un atributo, o bien la actividad electro-dérmica o bien la transpiración perinasal, y dicho atributo les falta en todas las muestras, por lo tanto no ha podido ser rellenado con la media de dicho atributo del propio sujeto.

Ahora tenemos varias opciones de cara a utilizar muchas de las técnicas de modelado que requieren que no haya valores ausentes, o bien eliminamos las filas con atributos con valores ausentes, con lo cual perderíamos más de 50.000 filas de datos (aproximadamente un 20% de los datos), o bien eliminamos los atributos con valores ausentes (perdemos variedad de información), o rellenamos los valores ausentes con la media del atributo para todo el conjunto (perdemos fidelidad).

En este trabajo se ha optado por la última opción, aunque se pierda fidelidad en algunos de los atributos, seguimos conservando toda la información que aportan los 50.000 ejemplos con datos ausentes en uno de estos dos atributos.

Por tanto para terminar de rellenar los datos, añadimos un último operador “replace missing values” al pre-procesamiento de los datos, que esta vez sustituirá los datos ausentes en los atributos Palm.EDA y Perinasal.Perspiration por la media de dichos atributos del conjunto total.

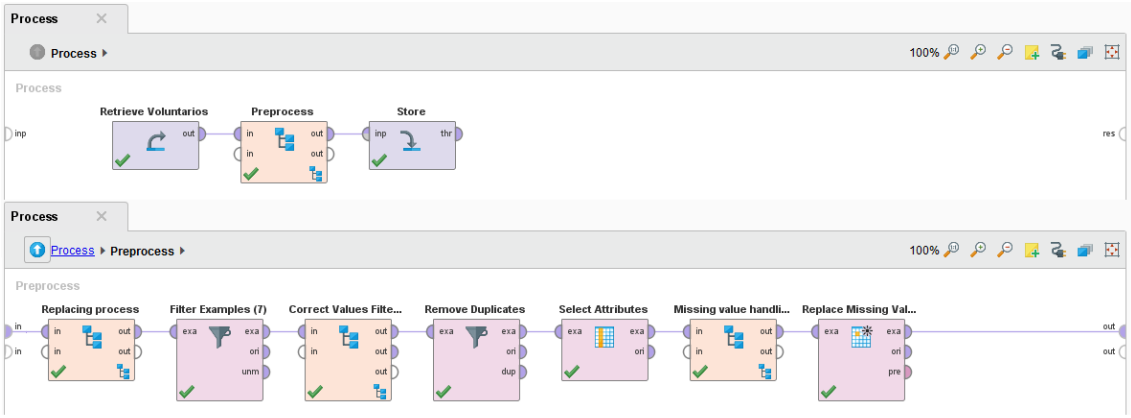


Fig. 21 Implementación final del preprocesamiento de los datos (elaboración propia, 2018)

Ahora si visualizamos la pestaña de estadísticas una vez más (figura 22), podemos comprobar que los problemas principales están resueltos y podemos comenzar a analizar los datos:

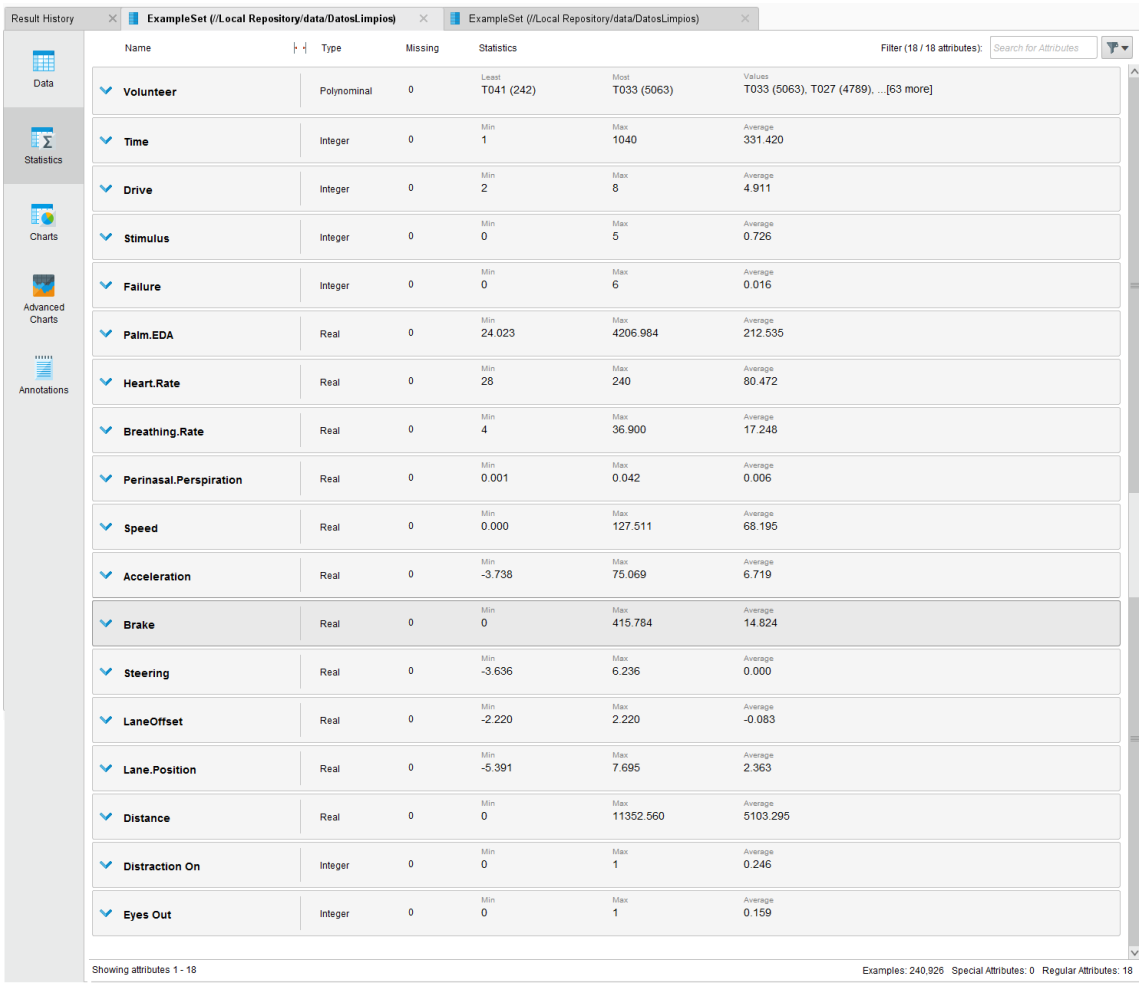


Fig. 22 Estadísticas de los datos tras el preprocesamiento (elaboración propia, 2018)

5. Modelado y análisis de los datos

En este capítulo se describirán en detalle los análisis aplicados a los diferentes conjuntos de datos, así como su implementación en Rapidminer, y posteriormente se explicarán las conclusiones obtenidas de los mismos.

5.1 Aplicación de las técnicas de minería de datos al conjunto de datos del simulador

5.1.1 Análisis de componentes principales (PCA)

Como hemos explicado anteriormente, los datos de los que disponemos no están etiquetados, así que centraremos nuestros análisis en técnicas descriptivas que nos ayuden a extraer la máxima información posible del conjunto de datos.

Primero intentaremos reducir la dimensión de nuestro conjunto de datos, puesto que tenemos casi 241.000 líneas de datos y más de 20 atributos distintos. Por tanto reducir el número de variables que evaluarán las técnicas de extracción de información puede simplificar la tarea, dado que en el aprendizaje no supervisado es especialmente difícil distinguir el conocimiento útil de un conjunto de datos grande.

Para ello se puede emplear el análisis de componentes principales (PCA), que es una técnica que busca la proyección de los datos en términos de mínimos cuadrados, según la cual los datos queden representados lo mejor posible [19]. El set de datos queda descrito en términos de nuevas variables o "componentes" no correlacionadas. Este tipo de análisis tiene sentido si existen altas correlaciones entre las variables, ya que esto es indicativo de que existe información redundante, y por tanto podría reducirse el conjunto a unos pocos factores que explicarían gran parte de la variabilidad total del set de datos.

La idea intuitiva es que las medidas biométricas, respiración, ritmo cardíaco y transpiración están relacionadas y podrían hallarse nuevas componentes que explicasen el conjunto con un número más reducido de atributos.

Comenzaremos realizando una matriz de correlaciones que analice cómo se relacionan los diferentes atributos de nuestro conjunto.

Para ello nos dirigiremos de nuevo a la pestaña proceso de Rapidminer y cargaremos nuestros datos previamente pre-procesados, realizaremos una unión de tablas con el conjunto de datos biográficos (conjunto correcto y sin valores ausentes) mediante el operador "join". Seleccionaremos como atributo de unión "volunteer" que es el identificador de voluntario, esto añadirá los atributos sexo, edad y grupo de edad (joven o anciano) a cada sujeto del set.

Después seleccionaremos los atributos del conjunto que queremos utilizar en el análisis mediante un operador "select attributes". Seleccionamos los atributos correspondientes a las medidas biométricas de los sujetos (ritmo cardíaco, respiración, transpiración perinasal, actividad electro-dérmica de las manos y el indicador que nos dice si el sujeto mira o no a la carretera "EyesOut"), las medidas del simulador (fuerza aplicada al freno, aceleración, velocidad, posición respecto a la línea, ángulo del volante, prueba de

conducción aplicada) y por último añadimos la edad de los datos biográficos (no se añade el sexo en este análisis por no ser un dato numérico, condición necesaria para el análisis).

Por último seleccionamos el operador "matriz de correlaciones" como se muestra a continuación (figura 23).

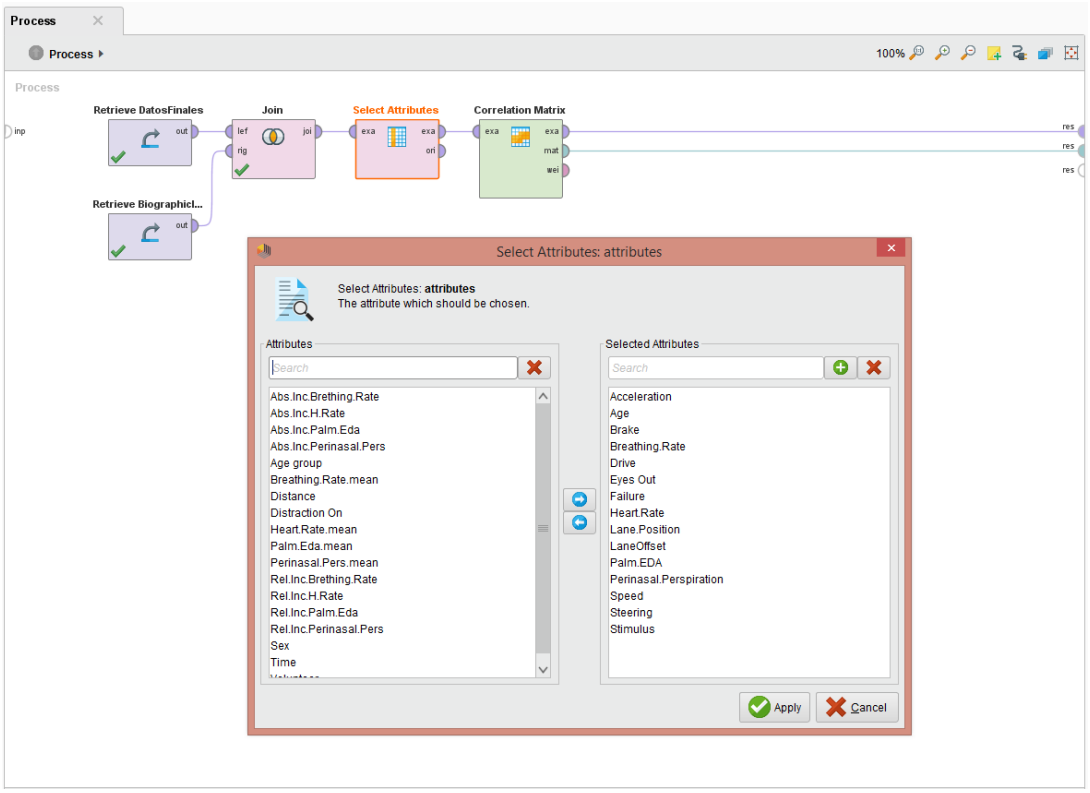


Fig. 23 Implementación de la matriz de correlaciones (elaboración propia, 2018)

Aplicamos el proceso y observamos la salida, que es la siguiente matriz de correlaciones (figura 24).

Attribut...	Drive	Stimulus	Failure	Palm.EDA	Heart.R...	Breathi...	Perinas...	Speed	Acceler...	Brake	Steering	LaneOff...	Lane.Po...	Eyes Out	Age
Drive	1	0.464	0.091	-0.070	-0.001	-0.058	0.090	0.024	-0.046	0.017	-0.000	0.098	-0.200	0.104	0.019
Stimulus	0.464	1	0.058	-0.009	0.027	0.103	0.099	0.025	-0.069	0.001	0.000	0.030	-0.203	0.197	-0.006
Failure	0.091	0.058	1	-0.004	0.011	-0.004	0.034	-0.066	-0.060	0.163	-0.002	-0.013	0.007	0.015	-0.002
Palm.EDA	-0.070	-0.009	-0.004	1	-0.082	0.080	-0.048	-0.052	0.052	-0.024	0.006	0.055	-0.010	0.057	0.252
Heart.Rate	-0.001	0.027	0.011	-0.082	1	-0.038	0.075	0.029	-0.004	-0.017	0.002	-0.026	0.004	-0.041	-0.280
Breathin...	-0.058	0.103	-0.004	0.080	-0.038	1	0.028	-0.019	0.002	-0.011	0.001	-0.017	0.071	0.115	-0.134
Perinasa...	0.090	0.099	0.034	-0.048	0.075	0.028	1	-0.032	-0.028	0.045	0.003	0.007	-0.031	0.016	0.163
Speed	0.024	0.025	-0.066	-0.052	0.029	-0.019	-0.032	1	0.097	-0.329	-0.006	0.016	0.065	-0.037	-0.056
Accelerat...	-0.046	-0.069	-0.060	0.052	-0.004	0.002	-0.028	0.097	1	-0.362	0.010	0.006	0.019	0.034	-0.010
Brake	0.017	0.001	0.163	-0.024	-0.017	-0.011	0.045	-0.329	-0.362	1	-0.003	-0.010	-0.014	-0.036	0.103
Steering	-0.000	0.000	-0.002	0.006	0.002	0.001	0.003	-0.006	0.010	-0.003	1	-0.122	0.089	-0.000	0.002
LaneOffs...	0.098	0.030	-0.013	0.055	-0.026	-0.017	0.007	0.016	0.006	-0.010	-0.122	1	-0.070	0.014	0.070
Lane.Po...	-0.200	-0.203	0.007	-0.010	0.004	0.071	-0.031	0.065	0.019	-0.014	0.089	-0.070	1	-0.043	-0.053
Eyes Out	0.104	0.197	0.015	0.057	-0.041	0.115	0.016	-0.037	0.034	-0.036	-0.000	0.014	-0.043	1	0.004
Age	0.019	-0.006	-0.002	0.252	-0.280	-0.134	0.163	-0.056	-0.010	0.103	0.002	0.070	-0.053	0.004	1

Fig. 24 Matriz de correlaciones (elaboración propia, 2018)

La correlación entre dos variables es una medida de la fuerza con que están relacionadas. Los coeficientes de la matriz variarán entre 0 y 1 y pueden tener signo positivo o negativo.

Las correlaciones positivas implican que las magnitudes están relacionadas de forma directa, de modo que cuando una aumenta, la otra también lo hace, y si una disminuye la otra también lo hará.

Las correlaciones negativas implican que las magnitudes están relacionadas de forma inversa, de modo que cuando una aumenta la otra disminuye y viceversa.

Así mismo es importante tener en cuenta el valor absoluto de dicha correlación, este indica la fuerza con la que están relacionadas ambas variables.

Como norma general se pueden usar las siguientes reglas como referencia [28]:

Correlación entre 0 y 0,4: Se considera que no existe correlación entre las variables, y aunque muestre una pequeña interacción entre atributos, no es estadísticamente relevante.

Correlación entre 0,4 y 0,6: Existe alguna correlación entre los atributos.

Correlación entre 0,6 y 1: Fuerte correlación entre los atributos, cuanto mayor es el número más fuerte es esta.

Nótese que correlación no implica causalidad. Sino solo que cuando una aumenta la otra también lo hace en el caso de la correlación positiva y que cuando una aumenta la otra disminuye en el caso de la correlación negativa.

Rapidminer utiliza un código de colores en la matriz para ayudarnos a ver rápidamente las magnitudes correlacionadas.

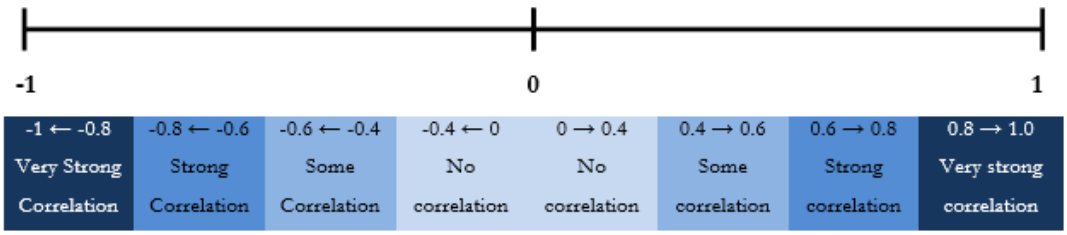


Fig. 25 Escala de color de la correlación en Rapidminer (Dr. Matthew A North- Data Mining for the Masses, 2012)[28]

Si nos fijamos ahora en la matriz de correlaciones de nuevo observamos que nuestras variables están muy poco correlacionadas, la mayoría de los coeficientes son inferiores a 0.1.

Las únicas excepciones son:

- Existe alguna correlación (0,464) entre la variable estímulo y el número de la prueba que se realiza, esto tiene lógica porque la variable estímulo es siempre 0 en las pruebas [1,3], y varía en el intervalo [1,5] para las pruebas [4,8].
- También se observa una pequeña correlación negativa entre la fuerza aplicada al freno y la velocidad (-0.329), y la fuerza aplicada al freno y la aceleración (-0,362). Esto tiene sentido porque aunque aquí sí que hay una relación de causalidad (al pisar el freno, disminuye la aceleración, y disminuye la velocidad). La aceleración y la velocidad aumentan sin que el freno se pise y su valor se mantiene en 0 cuando estas variables cambian.

- También se puede observar que hay una correlación que aunque no llega a ser significativa estadísticamente, se da entre la edad de los sujetos y su ritmo cardíaco (-0,280) y entre la edad y la transpiración perinasal (0,252), cuando aumenta la edad el ritmo cardíaco es menor, y su transpiración perinasal es mayor.
- Por último se observa una mínima relación entre la aplicación de distracciones y la pérdida de vista de la carretera por los sujetos (0.197). Aunque no parece estadísticamente relevante. Así como entre la posición respecto a la línea y el estímulo aplicado (-0.203).

Como hemos observado, no existen fuertes relaciones entre las variables analizadas, aunque hay que tener en cuenta que el ritmo cardíaco, respiración, y transpiración en reposo de cada persona varían según múltiples factores, su peso, su edad, actividad física, nivel de sedentarismo, etc. Por tanto las medidas absolutas de estos atributos variarán demasiado de un sujeto a otro y no serán comparables directamente. Ahora repetiremos el análisis utilizando los atributos Rel.Inc.Breathing.Rate, Rel.Inc.Heart.Rate, Rel.Inc.Perinasal.Perspiration y Rel.Inc.Palm.EDA que miden el incremento relativo de la respiración, ritmo cardíaco, transpiración perinasal y actividad electro-dérmica de las manos de los sujetos, respecto a la media de estas magnitudes en reposo.

La matriz de correlaciones obtenida con las nuevas variables es la siguiente (figura 26):

Attribut...	Drive	Stimulus	Failure	Speed	Acceler...	Brake	Steering	LaneOff...	Lane.Po...	Eyes Out	Rel.Inc....	Rel.Inc....	Rel.Inc....	Rel.Inc....	Age
Drive	1	0.464	0.091	0.024	-0.046	0.017	-0.000	0.098	-0.200	0.104	-0.089	0.011	-0.064	0.144	0.019
Stimulus	0.464	1	0.058	0.025	-0.069	0.001	0.000	0.030	-0.203	0.197	-0.056	0.068	0.122	0.188	-0.006
Failure	0.091	0.058	1	-0.066	-0.060	0.163	-0.002	-0.013	0.007	0.015	-0.005	0.019	-0.007	0.072	-0.002
Speed	0.024	0.025	-0.066	1	0.097	-0.329	-0.006	0.016	0.065	-0.037	-0.025	0.007	0.003	-0.068	-0.056
Accelerat...	-0.046	-0.069	-0.060	0.097	1	-0.362	0.010	0.006	0.019	0.034	-0.002	0.009	0.001	-0.016	-0.010
Brake	0.017	0.001	0.163	-0.329	-0.362	1	-0.003	-0.010	-0.014	-0.036	0.009	-0.015	0.006	0.058	0.103
Steering	-0.000	0.000	-0.002	-0.006	0.010	-0.003	1	-0.122	0.089	-0.000	0.016	0.000	0.001	0.005	0.002
LaneOffs...	0.098	0.030	-0.013	0.016	0.006	-0.010	-0.122	1	-0.070	0.014	0.014	-0.011	-0.030	0.016	0.070
Lane.Po...	-0.200	-0.203	0.007	0.065	0.019	-0.014	0.089	-0.070	1	-0.043	0.017	-0.013	0.072	-0.066	-0.053
Eyes Out	0.104	0.197	0.015	-0.037	0.034	-0.036	-0.000	0.014	-0.043	1	0.031	0.029	0.031	0.088	0.004
Rel.Inc.P...	-0.089	-0.056	-0.005	-0.025	-0.002	0.009	0.016	0.014	0.017	0.031	1	-0.001	-0.013	0.009	-0.050
Rel.Inc.H...	0.011	0.068	0.019	0.007	0.009	-0.015	0.000	-0.011	-0.013	0.029	-0.001	1	-0.019	0.105	-0.052
Rel.Inc.B...	-0.064	0.122	-0.007	0.003	0.001	0.006	0.001	-0.030	0.072	0.031	-0.013	-0.019	1	0.015	0.028
Rel.Inc.P...	0.144	0.188	0.072	-0.068	-0.016	0.058	0.005	0.016	-0.066	0.088	0.009	0.105	0.015	1	0.057
Age	0.019	-0.006	-0.002	-0.056	-0.010	0.103	0.002	0.070	-0.053	0.004	-0.050	-0.052	0.028	0.057	1

Fig. 26 Matriz de correlaciones tras la adición de los nuevos atributos (elaboración propia, 2018)

Como podemos observar, en la nueva matriz se mantienen las relaciones entre los atributos que vimos anteriormente, y entre los nuevos atributos tampoco se observan fuertes relaciones ni entre ellos ni con las otras variables. Por tanto podemos suponer que el análisis de componentes principales reducirá poco el número de variables totales.

Ahora procederemos a realizar el análisis de componentes principales en Rapidminer. Primero seleccionamos los atributos principales (el dataset tiene algunos atributos que se utilizan solo para calcular otros, que no queremos eliminar, pero tampoco los usamos directamente), estos son: Drive, Stimulus, Failure, Palm.EDA, Heart.Rate, Breathing.Rate, Perinasal.Perspiration, Speed, Acceleration, Brake, Steering, LaneOffset, LanePosition, EyesOut, y Age.

Ahora debemos normalizar los atributos de nuestro dataset. Utilizando el operador normalize, seleccionamos todos los atributos y los normalizamos mediante Transformada-Z, de este modo todos estarán en la misma escala [30].

Ahora conectamos el operador PCA, seleccionamos mantener varianza, con un umbral de varianza del 95%, de manera que solo utilizará componentes hasta completar un 95% de la varianza total del conjunto de datos. Pudiendo perder como máximo un 5% de la varianza total en el proceso. La implementación en Rapidminer se muestra a continuación (figura 27):

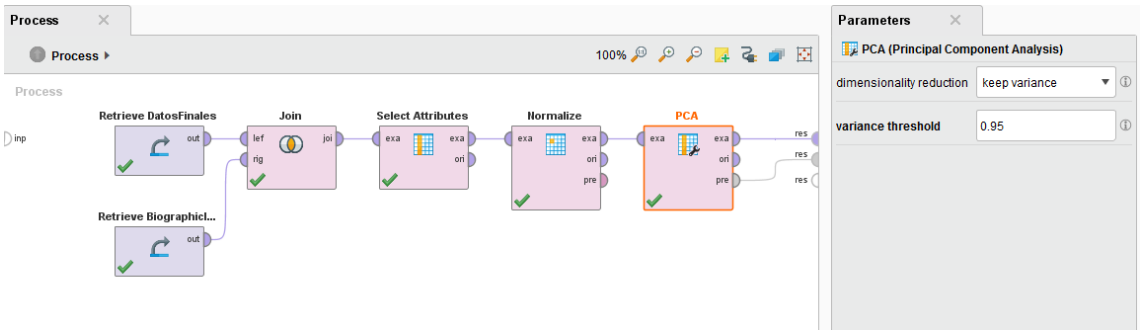


Fig. 27 Implementación del análisis de componentes principales (elaboración propia, 2018)

Este análisis nos proporciona dos salidas:

- La primera es la salida del conjunto, en la que cada entrada contendrá ahora un valor para cada una de las componentes principales (nuevos atributos incorrelados). Estos serán los datos que usaremos si reducimos el conjunto aplicando esta técnica.
- La segunda será el premodelado que nos proporcionará dos ficheros, eigenvalues y eigenvectors:
  - Eigenvalues es una matriz que contiene en primer lugar los autovalores de cada componente principal, en segundo lugar la proporción de varianza explicada por cada componente principal y por último la varianza acumulada. Las componentes se encontrarán ordenadas de mayor a menor varianza explicada, para facilitar la elección del número de componentes principales que usaríamos para representar nuestro conjunto según la proporción de varianza total que queramos conservar. No hay una regla para decidir cuantas componentes elegir, pero por norma general se trata de reducir el conjunto lo máximo posible, perdiendo la menor varianza posible.
  - Eigenvectors: esta matriz contiene los coeficientes factoriales de las variables (matriz de correlaciones de cada componente principal con cada una de las variables originales). Esto nos ayudará a interpretar los nuevos atributos.

Observamos la salida de la matriz de autovalores (figura 28).

Result History | PCA (PCA) | ExampleSet (PCA)

Eigenvalues | Eigenvectors | Cumulative Variance | Annotations

Component	Standard Deviation	Proportion of Variance	Cumulative Variance
PC 1	1.333	0.118	0.118
PC 2	1.261	0.106	0.224
PC 3	1.202	0.096	0.321
PC 4	1.103	0.081	0.402
PC 5	1.054	0.074	0.476
PC 6	1.012	0.068	0.544
PC 7	0.985	0.065	0.609
PC 8	0.978	0.064	0.673
PC 9	0.951	0.060	0.733
PC 10	0.931	0.058	0.791
PC 11	0.913	0.056	0.846
PC 12	0.860	0.049	0.895
PC 13	0.754	0.038	0.933
PC 14	0.716	0.034	0.968
PC 15	0.698	0.032	1.000

Fig. 28 Autovalores y proporción de varianza del análisis de componentes principales (elaboración propia, 2018)

Como ya suponíamos previamente por los resultados de la matriz de correlaciones, de los 15 atributos principales que tiene nuestro conjunto se necesitarían 14 componentes para obtener una varianza del 96,8%, superando así el valor recomendado del operador de Rapidminer, y si usásemos solo 13 nuestro conjunto mantendría un 93,3% de la varianza. Si miramos la proporción de varianza de cada atributo vemos que la máxima proporción de varianza es menor al 12% de la varianza total del conjunto y el atributo de menor proporción de varianza tiene un 3,2%. Por lo que solo sería recomendable reducir el conjunto a 14 atributos.

Si ahora observamos la tabla EigenValues podremos identificar cómo se relacionan las componentes principales que hemos creado con cada variable, hay 3 factores que facilitan la interpretación de la tabla:

- Que los coeficientes sean próximos a 1.
- Que cada componente tenga solo coeficientes elevados con una variable.
- No deben existir factores con coeficientes similares.

Result History | PCA (PCA) | ExampleSet (PCA)

Eigenvalues | Eigenvectors | Cumulative Variance | Annotations

Attribute	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6	PC 7	PC 8	PC 9	PC 10	PC 11	PC 12	PC 13	PC 14	PC 15
Drive	0.504	-0.298	-0.037	-0.110	0.137	0.164	-0.114	-0.036	0.087	0.033	-0.090	0.370	-0.058	-0.359	-0.546
Stimulus	0.508	-0.332	-0.042	0.147	0.071	0.076	-0.003	-0.093	0.089	-0.141	-0.009	0.251	0.071	0.381	0.590
Failure	0.202	0.150	-0.157	0.074	0.072	0.020	-0.705	0.532	-0.151	-0.107	-0.073	-0.244	-0.085	0.123	-0.019
Palm.EDA	0.010	0.119	0.488	0.246	-0.098	-0.067	0.030	0.295	0.477	-0.391	0.189	0.113	-0.246	-0.291	0.106
Heart.Rate	-0.050	-0.163	-0.464	-0.099	-0.027	-0.380	0.191	0.364	0.412	-0.076	0.282	0.048	0.287	0.213	-0.214
Breathing.R...	0.010	-0.079	-0.034	0.644	-0.271	-0.261	-0.056	-0.251	0.007	-0.231	-0.422	-0.107	0.215	0.068	-0.276
Perinasal.P...	0.214	0.014	0.017	-0.056	0.397	-0.774	0.068	-0.097	-0.177	0.041	-0.059	-0.119	-0.255	-0.216	0.131
Speed	-0.198	-0.422	0.061	-0.155	0.156	0.014	-0.396	-0.338	0.108	-0.287	0.276	-0.331	0.316	-0.266	0.095
Acceleration	-0.256	-0.357	0.271	0.027	0.075	-0.084	0.042	0.482	-0.231	0.188	-0.319	0.238	0.402	-0.216	0.175
Brake	0.287	0.544	-0.210	0.025	-0.072	0.024	0.011	-0.056	0.048	0.047	0.030	0.047	0.544	-0.452	0.245
Steering	-0.062	0.031	-0.064	0.273	0.621	0.260	0.109	0.029	0.465	0.285	-0.254	-0.297	0.007	0.003	0.014
LaneOffset	0.134	-0.068	0.219	-0.275	-0.439	-0.204	-0.257	-0.087	0.439	0.519	-0.218	-0.159	-0.013	0.051	0.086
Lane.Positi...	-0.340	0.136	-0.087	0.265	0.158	-0.154	-0.446	-0.213	0.098	0.278	0.218	0.595	-0.051	0.058	0.032
Eyes Out	0.217	-0.190	0.118	0.457	-0.119	0.029	0.092	0.095	-0.215	0.452	0.580	-0.252	0.029	-0.076	-0.070
Age	0.165	0.263	0.568	-0.133	0.276	-0.098	-0.028	-0.059	-0.038	-0.040	0.119	0.026	0.411	0.443	-0.299

Fig. 29 Matriz de coeficientes factoriales de los atributos (elaboración propia, 2018)

Si tratamos de interpretar cada componente, observamos que la componente 1, PC1 tiene coeficientes elevados y positivos con las variables Drive y Stimulus, que son bajas en el

resto de las componentes a excepción de la componente 15, que no se usaría. El resto de las componentes tienen una correlación baja por lo que relacionaríamos esta variable con las pruebas de conducción en las cuales se distrae al sujeto.

En la componente 2, PC2 observamos que la correlación positiva es alta con la fuerza de frenado y negativa y moderadamente alta con la velocidad y la aceleración, por tanto esta componente se relaciona directamente con la magnitud de frenado.

En la componente3, PC3 observamos una correlación alta y positiva con la actividad electro-dérmica de las manos y con la edad, y una correlación alta y negativa con el ritmo cardíaco, por tanto estará relacionada con los sujetos de mayor edad (grupo anciano), con una transpiración perinasal alta y menor ritmo cardíaco.

Podríamos continuar relacionando las componentes principales con las magnitudes a las que hacen referencia, pero como se puede observar este método no es nada práctico para datasets que no tienen alta correlación (y por tanto información redundante), porque apenas podríamos reducir un solo atributo perdiendo un 3,2% de la varianza total del conjunto, y además incrementamos enormemente la dificultad de entender qué información nos proporciona cada atributo, por tanto, no usaremos dicho método en nuestro conjunto de datos.

5.1.2 Algoritmo de agrupamiento: K-means

K-means es un algoritmo de agrupamiento que forma k grupos de objetos basándose en la similitud de sus características. El agrupamiento se realiza minimizando la suma de distancias entre cada objeto y el centroide de su grupo o cluster. Se suele usar la distancia cuadrática [18].

El algoritmo consta de tres pasos:

1. Primero se elige el número de grupos a realizar, una vez determinado este parámetro k, se establecen k centroides en el espacio de los datos escogiéndolos aleatoriamente.
2. Asignación de objetos a los centroides: cada objeto de los datos es asignado a su centroide más cercano según la distancia o métrica elegida.
3. Actualización de los centroides: se actualiza la posición del centroide de cada grupo tomando como nuevo centroide la posición del promedio de los objetos pertenecientes a dicho grupo.

Los pasos 2 y 3 se repiten hasta que los centroides se mueven una magnitud menor a una distancia umbral.

Las distancias se pueden calcular de múltiples maneras dependiendo del problema, aunque estas son las distancias más comunes para valores continuos:

Se necesita normalizar los datos previamente [29]:

- Distancia Euclídea: $\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$
- Distancia de Manhattan: $\sum_{i=1}^{n}|x_i - y_i|$

- Distancia de Chebychev: $max_{i=1\ldots n}|x_i - y_i|$

No es necesario normalizar los datos:

- Distancia del coseno: cada ejemplo es un vector y la distancia es el coseno del ángulo que forman.

5.1.2.1 Primera aplicación del agrupamiento K-means

Ahora procederemos a realizar el agrupamiento K-means en Rapidminer, comparando los resultados obtenidos según el valor de k, utilizando la métrica euclídea. Los grupos obtenidos dependerán del valor de k y del punto de convergencia del algoritmo, que no será un mínimo global, sino local. Es trabajo del analista determinar si los grupos obtenidos tienen sentido o si son simples agrupaciones de características, también es difícil definir el concepto de error en las técnicas no supervisadas, pero posteriormente utilizaremos algunas técnicas de validación para comprobar la eficacia de los agrupamientos realizados.

Cabe destacar que para poder realizar un análisis de agrupamiento no pueden existir valores ausentes en el conjunto, además son altamente sensibles a la presencia de "outliers" (valores muy por encima o por debajo de los valores normales).

Primero volvemos a seleccionar el conjunto de datos ya preprocesados, que no tienen valores ausentes, y cuyas medidas han sido filtradas a los valores mínimos y máximos que detectan los sensores. Por tanto la presencia de "outliers" en nuestro conjunto es mínima, y además no queremos eliminar datos dentro del rango de detección de los sensores, porque puede aportar información útil. Unimos dicho set con los datos biográficos de los usuarios mediante un operador "unión" como en el caso anterior añadiendo los atributos sexo, edad y grupo de edad.

Seleccionamos los atributos que van a participar en el análisis, en este caso utilizamos las medidas del simulador (fuerza del freno, velocidad, aceleración, posición respecto a la línea, etc), las magnitudes que miden el aumento de las medidas biométricas respecto a la media en estado de reposo del sujeto y el sexo y la edad de cada sujeto que acabamos de añadir.

Normalizamos de nuevo los datos mediante transformada-z con el operador "normalize", puesto que es necesario para que los resultados del agrupamiento sean correctos, y conectamos el operador "K-means" con métrica de distancia euclídea y dos grupos. A continuación puede verse el diagrama de bloques en Rapidminer (figura 30).

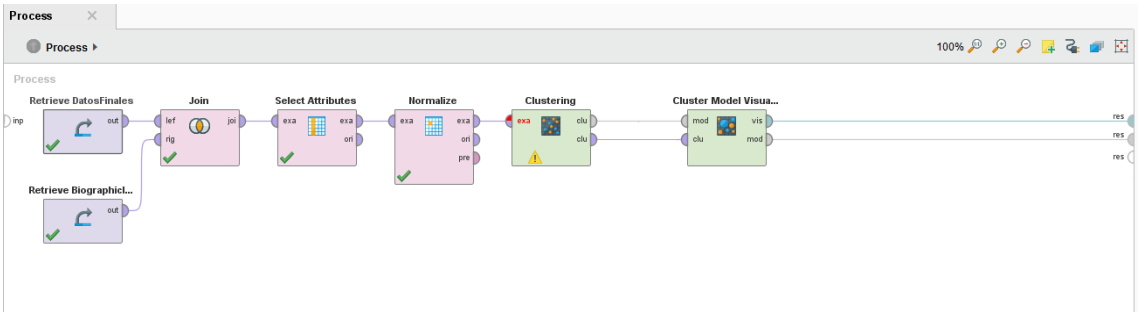


Fig. 30 Implementación del agrupamiento K-means (elaboración propia, 2018)

Añadimos el operador "cluster model visualizer" para poder observar con mayor claridad los agrupamientos que vamos a realizar.

Seleccionamos dos grupos y pulsamos "Start" para obtener la salida del agrupamiento. Estos son los resultados obtenidos:

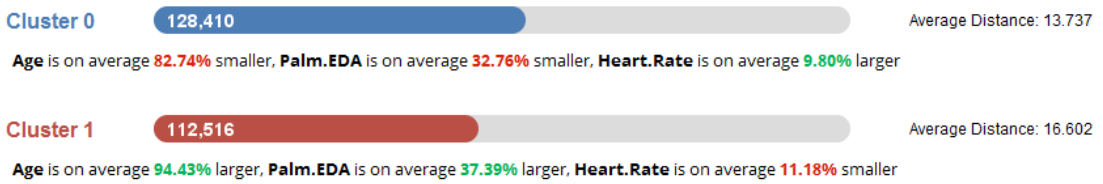


Fig. 31 Vista preliminar de la primera aplicación de agrupamiento K-means (elaboración propia, 2018)

Attribute	cluster_0	cluster_1
Stimulus	0.001	-0.001
Palm.EDA	-0.255	0.291
Heart.Rate	0.243	-0.277
Breathing.Rate	0.134	-0.153
Perinasal.Perspiration	-0.165	0.189
Speed	0.054	-0.062
Acceleration	0.008	-0.009
Steering	-0.002	0.002
Lane.Position	0.051	-0.058
Eyes Out	-0.008	0.009
Rel.Inc.Palm.Eda	0.051	-0.059
Rel.Inc.H.Rate	0.056	-0.064
Rel.Inc.Brething.Rate	-0.015	0.018
Rel.Inc.Perinasal.Pers	-0.049	0.056
Age	-0.919	1.049
Sex	1.526	1.451

Tabla 5 Tabla de centroides de la primera aplicación de agrupamiento K-means (elaboración propia, 2018)

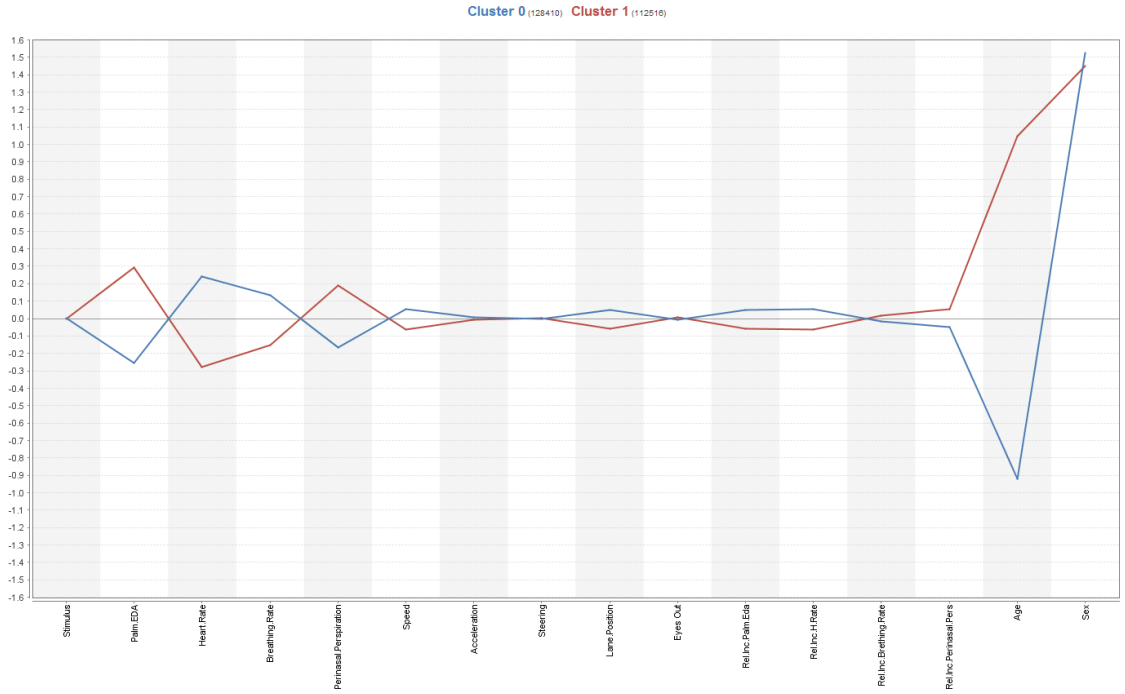


Fig. 32 Gráfica de centroides de la primera aplicación de agrupamiento K-means (elaboración propia, 2018)

Como puede observarse, la edad es un factor muy significativo en este agrupamiento, el primer grupo está formado por personas ancianas (con edad superior a 60 años), cuya actividad electro-dérmica es superior a la del otro grupo y su ritmo cardíaco más bajo. El segundo grupo está formado por individuos jóvenes con mayor ritmo cardíaco y menor actividad electro-dérmica en las manos. Además se puede observar que la transpiración perinasal también es significativamente mayor en el grupo de los ancianos. Sin embargo ambos grupos tienen aproximadamente el mismo número de hombres y de mujeres.

Al ser trabajo del analista decidir qué atributos utilizar a la hora de realizar los análisis para obtener información útil de ellos, se va a realizar la siguiente agrupación de tres grupos utilizando las mismas variables y posteriormente se elegirá si se retiran algunas para poder ver más claramente los efectos de las variables más interesantes de cara a buscar factores de riesgo (o distracciones) que causen cambios notables en la actividad de conducción de los sujetos. A continuación se muestra la gráfica de centroides realizada con los mismos atributos para el agrupamiento de tres grupos.

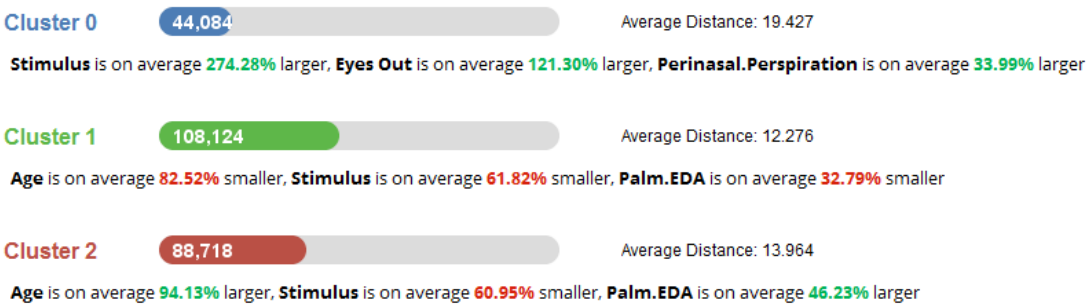


Fig. 33 Vista preliminar de la segunda aplicación de agrupamiento K-means (elaboración propia, 2018)

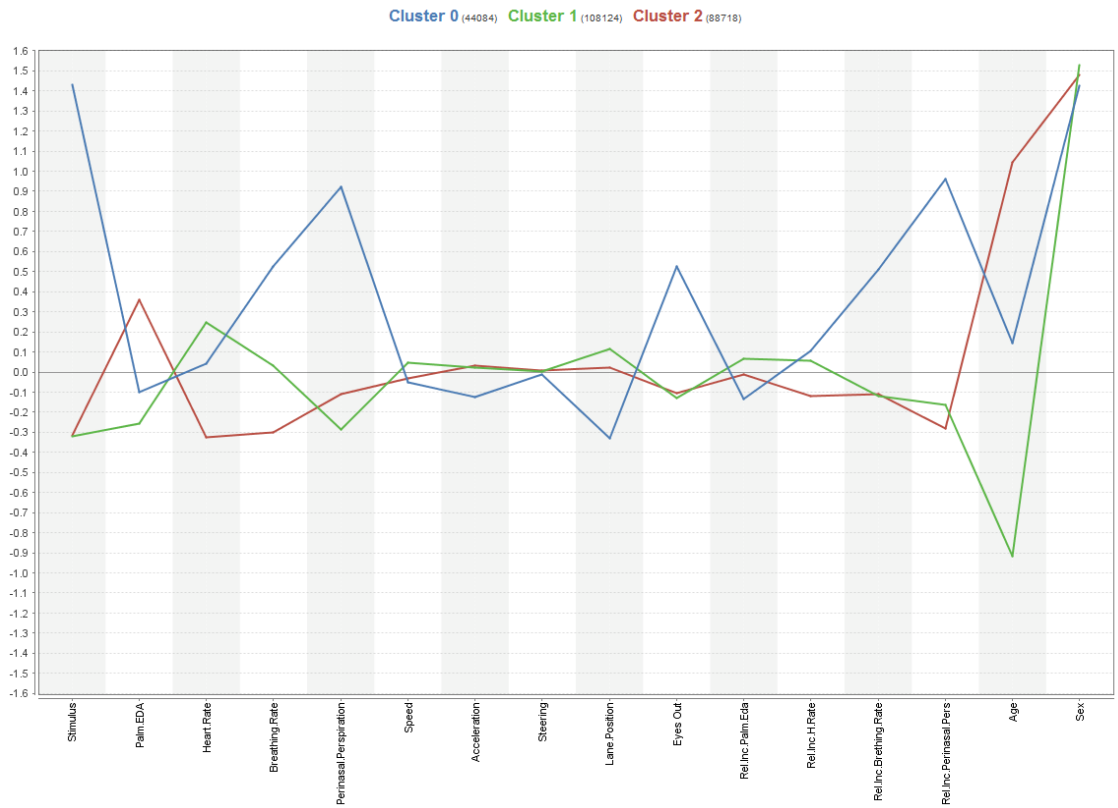


Fig. 34 Gráfica de centroides de la segunda aplicación de agrupamiento K-means (elaboración propia, 2018)

Como se puede observar, la edad sigue siendo el principal factor de separación de los grupos. Los grupos 1 y 2 son muy parecidos a los que se obtuvieron en el primer agrupamiento, uno formado por jóvenes con menor actividad electro-dérmica de las manos y transpiración perinasal, pero mayor ritmo cardíaco, y otro grupo de ancianos, con menor ritmo cardíaco, pero mayor actividad electro-dérmica de las manos y transpiración perinasal, el tercer grupo está formado por las muestras en las que los sujetos están sometidos a una distracción, y el número de veces que pierden la vista de la carretera es muy superior al resto de grupos (a los cuales no se les aplica distracciones), así como la transpiración perinasal es muy alta en este grupo. De nuevo puede observarse que los grupos obtenidos tienen una cantidad muy aproximada de hombres y mujeres, no siendo relevante el sexo en esta agrupación.

Se puede inferir con este análisis preliminar que las magnitudes biométricas de los sujetos se ven afectadas por su edad, y los grupos formados por similitud de características tienden a priorizar la edad de los sujetos a la hora de elegir grupos. Aunque esto es de esperar, no es de interés seguir incluyendo esta variable en las agrupaciones, puesto que queremos observar la influencia que causan las distracciones sobre los sujetos de las pruebas y esta variable "ocluye" las variaciones que se producen en los grupos causadas por los factores que queremos investigar.

De la misma manera y tras la realización de más agrupaciones se ha eliminado la variable que mide la fuerza del freno de este análisis por el mismo motivo, debido a que los valores de dicha variable cambiaban súbitamente de cero a cientos de newtons, por tanto tomaba gran relevancia en las agrupaciones, que consideraba muy distintos a un grupo en el que

los usuarios no estuviesen pisando el freno y otro en el que sí lo hiciesen, resultando agrupaciones condicionadas enormemente por esta variable. Por eso no se encuentra dicha variable en las gráficas y tablas, aunque no se incluya el análisis justificatorio de su eliminación, similar al de la variable edad, por su carácter redundante. Dado que los grupos obtenidos por la edad aportan información en alguna medida, pero los grupos obtenidos basados en la variable freno solo distinguen los momentos en los que los sujetos están frenando.

5.1.2.2 K-means con K=2

Una vez seleccionadas las variables que mejor van a permitir estudiar las distracciones al volante, se realizará de nuevo el agrupamiento con dos grupos y se procederán a estudiar sus resultados:

Como se esperaba, en este caso si se obtiene un papel más relevante de los estímulos aplicados y las magnitudes biométricas.

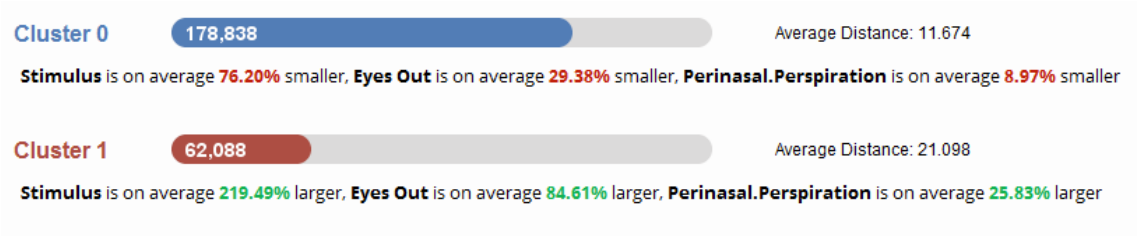


Fig. 35 Vista preliminar del agrupamiento K-means para k=2 (elaboración propia, 2018)

Attribute	cluster_0	cluster_1
Stimulus	-0.397	1.143
Palm.EDA	0.026	-0.074
Heart.Rate	-0.165	0.476
Breathing.Rate	-0.121	0.349
Perinasal.Perspiration	-0.244	0.702
Speed	-0.003	0.009
Acceleration	0.029	-0.082
Steering	0.004	-0.010
Lane.Position	0.103	-0.297
Eyes Out	-0.128	0.368
Rel.Inc.Palm.Eda	0.041	-0.119
Rel.Inc.H.Rate	-0.164	0.472
Rel.Inc.Brething.Rate	-0.123	0.353
Rel.Inc.Perinasal.Pers	-0.259	0.745
Sex	1.514	1.425

Tabla 6 Tabla de centroides para K-means k=2 (elaboración propia, 2018)

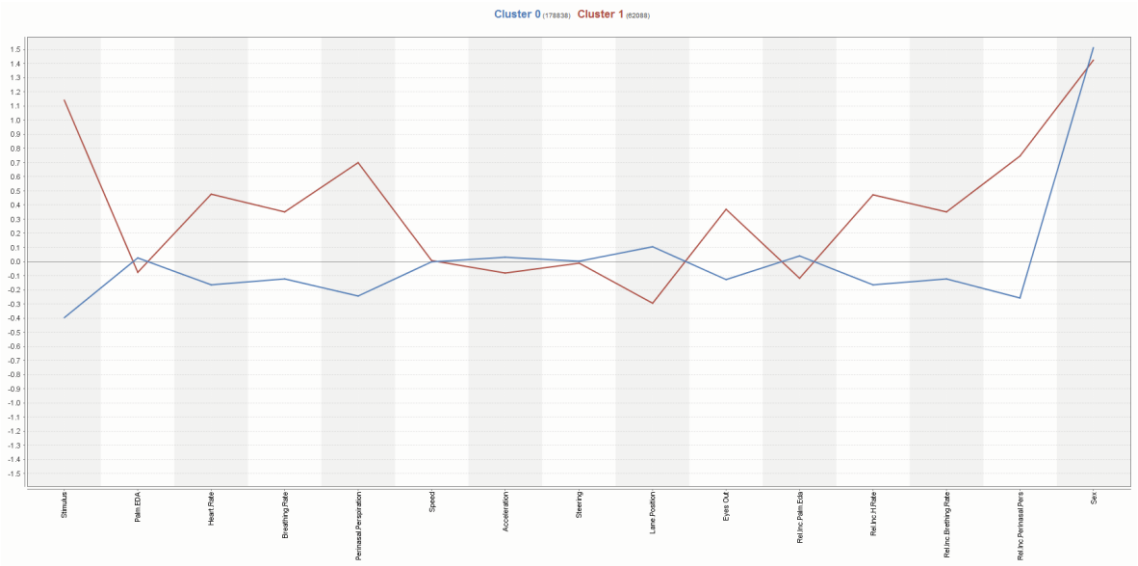


Fig. 36 Gráfica de los centroides para K-means con k=2 (elaboración propia, 2018)

Como se puede observar en la vista preliminar de la salida, se han creado dos grupos distintos con una cantidad desequilibrada de objetos clasificados en cada uno de ellos.

- El primer grupo, grupo 0, tiene 178.838 muestras clasificadas entre las cuales el atributo estímulo toma en su mayoría valores bajos (0 en casi todos ellos, no se aplica estímulo/distracción). Además la cantidad de muestras correspondientes a la pérdida de vista de la carretera es significativamente menor que en el otro grupo. La tercera característica más significativa señalada por el programa es que la transpiración perinasal es también menor que en el otro grupo. Aparte de estas características observamos que el primer grupo tiene en general un ritmo cardíaco y respiración (así como las variables que miden el incremento de ritmo cardíaco, ritmo respiratorio y transpiración perinasal respecto a las medias del propio sujeto) muy inferiores a los valores del otro grupo.
- El segundo grupo, grupo 1, tiene solo 62.088 muestras clasificadas. En dichas muestras el atributo estímulo toma valores superiores a 1 (eso quiere decir que se está aplicando una distracción). En este grupo la cantidad de veces que los sujetos pierden de vista la carretera es muy superior al grupo 0 y su transpiración perinasal es superior al otro grupo. De manera contraria al primer caso los valores del ritmo cardíaco, respiración y transpiración perinasal, así como las variables que miden el incremento de dichas magnitudes para cada sujeto son muy superiores al grupo 0.

Por tanto se puede suponer que el algoritmo ha diferenciado los dos grupos principalmente por la aplicación de las distracciones. El grupo 0 está formado por individuos a los que no se les está aplicando distracción los cuales pierden menos veces la vista de la carretera y sus medidas biométricas se encuentran en un nivel muy inferior a los del otro grupo porque no manifiestan ningún estrés o ansiedad. El grupo 1 sin embargo se compone de aquellas muestras de los individuos a los que se les está aplicando

una distracción y como consecuencia pierden más veces la atención de la carretera y sus medidas biométricas son más elevadas debido el estrés.

5.1.2.3 K-means con K=3

Volveremos a realizar ahora el agrupamiento K-means con 3 grupos y analizaremos la nueva agrupación:

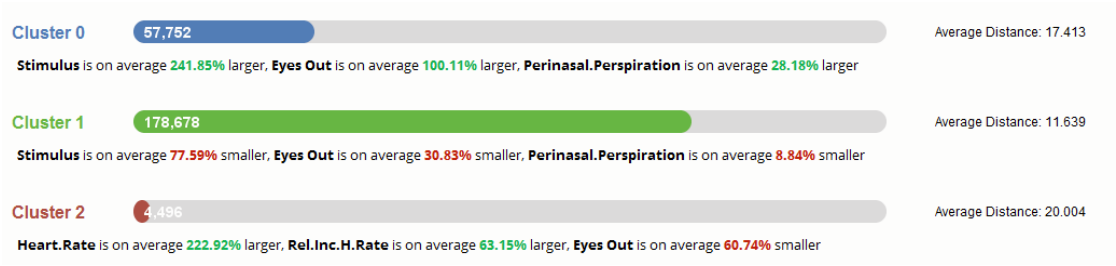


Fig. 37 Vista preliminar del agrupamiento K-means para k=3 (elaboración propia, 2018)

Attribute	cluster_0	cluster_1	cluster_2
Stimulus	1.260	-0.404	-0.119
Palm.EDA	-0.067	0.023	-0.045
Heart.Rate	0.040	-0.152	5.529
Breathing.Rate	0.440	-0.127	-0.588
Perinasal.Perspiration	0.765	-0.240	-0.291
Speed	-0.011	-0.001	0.172
Acceleration	-0.094	0.031	-0.021
Steering	-0.011	0.003	0.002
Lane.Position	-0.324	0.104	0.019
Eyes Out	0.436	-0.134	-0.264
Rel.Inc.Palm.Eda	-0.102	0.043	-0.403
Rel.Inc.H.Rate	0.113	-0.146	4.346
Rel.Inc.Brething.Rate	0.407	-0.126	-0.210
Rel.Inc.Perinasal.Pers	0.797	-0.256	-0.080
Sex	1.441	1.513	1.273

Tabla 7 Tabla de centroides para K-means k=3 (elaboración propia, 2018)

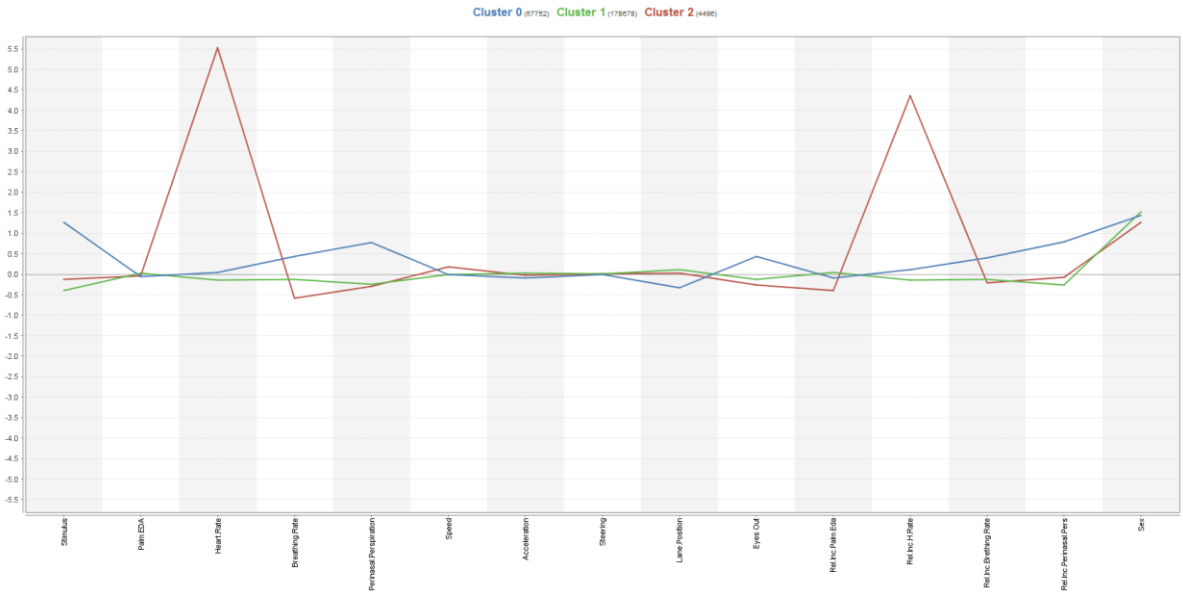


Fig. 38 Gráfica de los centroides para K-means con k=3 (elaboración propia, 2018)

En esta agrupación los tres grupos formados son muy similares al caso anterior:

- El grupo 0 está formado por 57.752 muestras, entre las cuales la mayoría corresponden a valores mayores a 1 de la variable estímulo, por tanto este grupo está formado por una mayoría de muestras donde a los sujetos se les aplica una distracción, y el porcentaje de pérdida de visión de la carretera es mucho más alto que en los otros grupos, además igual que en el caso anterior, los sujetos de este grupo tienen el ritmo cardíaco, respiración y transpiración perinasal, así como las variables que miden el incremento relativo de estas variables respecto a su media en reposo superiores a los otros grupos (excepto al último que es una agrupación con ritmo cardíaco alto, pero no respiración o transpiración).
- El grupo 1 está formado por 178.678 muestras contrarias al primer grupo, apenas se les aplican distracciones y el porcentaje de pérdida de vista de la carretera es significativamente menor que en el primer grupo. La transpiración, el ritmo cardíaco y la respiración, así como los atributos que miden el incremento de los mismos son menores al primer grupo.
- El grupo 2 está formado por 4496 muestras que tienen un ritmo cardíaco e incremento del ritmo cardíaco superior a los otros grupos. No ocurre lo mismo con el resto de medidas biométricas que son incluso inferiores a los valores de los otros grupos. La pérdida de visión de la carretera también es significativamente baja en este grupo. La variable estímulo también toma valores bajos (0) en su mayoría. Este grupo debe estar constituido por muestras de los sujetos a los que se les ha dejado de aplicar una distracción hace breves instantes, por tanto mantiene el incremento de ritmo cardíaco, pero al no aumentar la respiración ni la transpiración el algoritmo ha juzgado las muestras suficientemente diferentes para crear un grupo por sí mismas.

Como podemos observar los grupos son similares a los del agrupamiento k=2, pero aproximadamente 4000 muestras del grupo 1 de ese análisis han formado un nuevo grupo

(el grupo 2). No se ha podido obtener de este análisis más información que la que se obtuvo con k=2.

5.1.2.4 K-means con K=4

Se procederá ahora a realizar el agrupamiento con cuatro grupos:

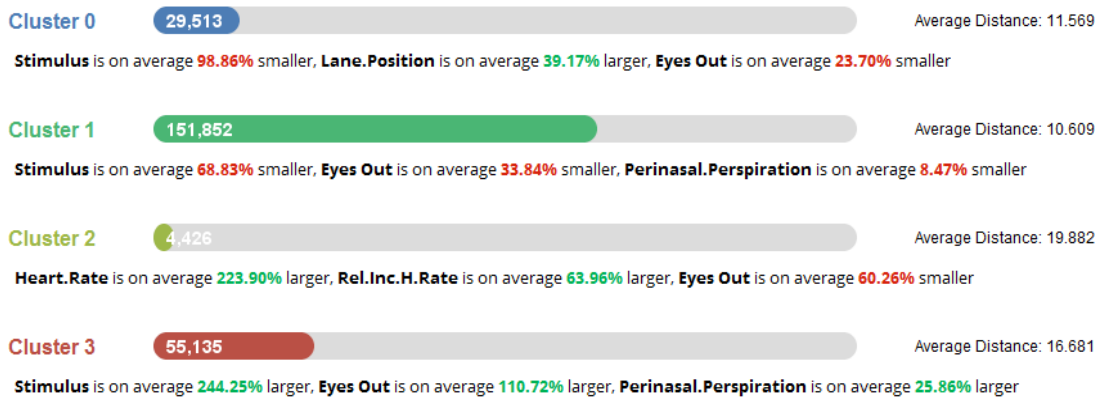


Fig. 39 Vista preliminar del agrupamiento K-means para k=4 (elaboración propia, 2018)

Attribute	cluster_0	cluster_1	cluster_2	cluster_3
Stimulus	-0.515	-0.358	-0.114	1.272
Palm.EDA	0.010	0.026	-0.044	-0.074
Heart.Rate	-0.095	-0.158	5.554	0.040
Breathing.Rate	0.196	-0.215	-0.591	0.535
Perinasal.Perspiration	-0.084	-0.230	-0.292	0.702
Speed	0.181	-0.036	0.173	-0.013
Acceleration	0.058	0.024	-0.016	-0.095
Steering	0.130	-0.021	0.001	-0.011
Lane.Position	2.549	-0.355	-0.013	-0.386
Eyes Out	-0.103	-0.147	-0.262	0.482
Rel.Inc.Palm.Eda	0.062	0.037	-0.403	-0.102
Rel.Inc.H.Rate	-0.106	-0.151	4.402	0.120
Rel.Inc.Brething.Rate	0.185	-0.209	-0.223	0.494
Rel.Inc.Perinasal.Pers	-0.174	-0.244	-0.078	0.773
Sex	1.498	1.511	1.272	1.448

Tabla 8 Tabla de centroides para K-means k=4 (elaboración propia, 2018)

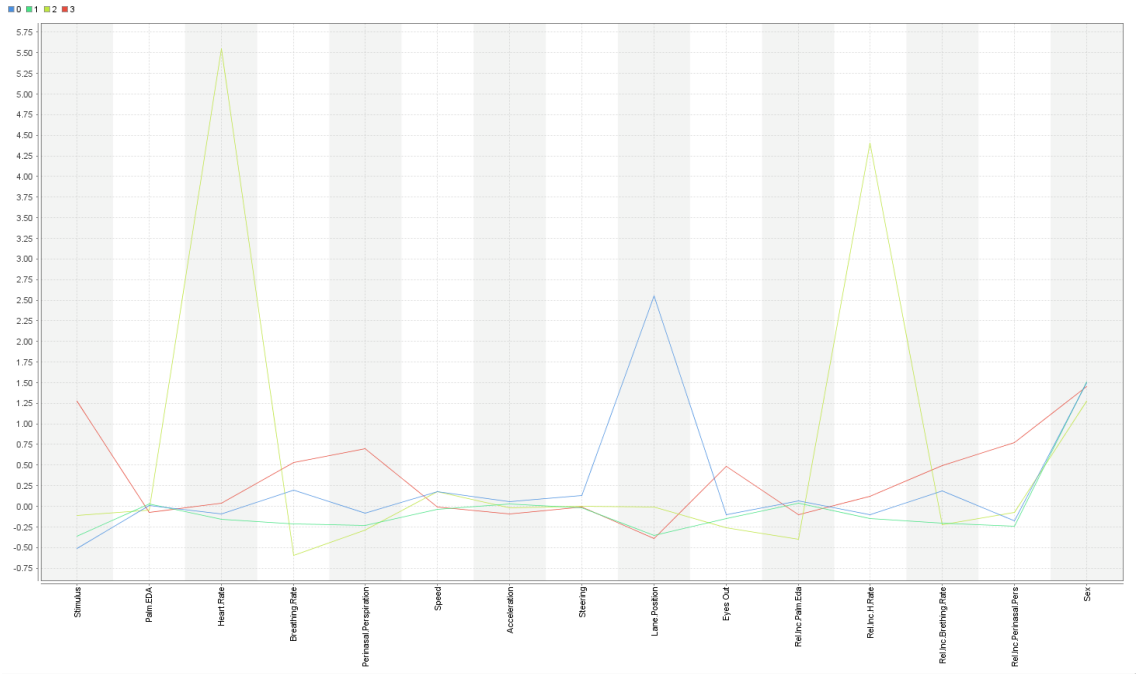


Fig. 40 Gráfica de los centroides para K-means con k=4 (elaboración propia, 2018)

El agrupamiento de cuatro grupos parece que es de nuevo una agrupación de características, sin representar adecuadamente grupos homogéneos de comportamientos.

- El primer grupo, grupo 0, es un grupo relativamente pequeño que estaría formado en su mayoría por personas a las que no se está distrayendo, el número de veces que pierden la vista de la carretera es menor que la media, y la variable Lane.Position nos indica que en dichos instantes el coche estaba situado más a la izquierda del límite del carril que el resto de grupos.
- El segundo grupo, grupo 1, está formado igualmente por sujetos que no están siendo distraídos, en la mayoría de muestras, la pérdida de visión de la carretera es baja, y la transpiración perinasal es baja (al igual que el ritmo cardíaco y la respiración). Este grupo es el grupo principal de individuos a los que no se les está aplicando una distracción.
- En el tercer grupo, grupo 2, se encuentran agrupadas las muestras en las que los sujetos tienen el ritmo cardíaco acelerado, aunque no se les está aplicando una distracción, la pérdida de visión de la carretera es baja.
- El cuarto grupo, grupo 3, agrupa los sujetos a los que se les está distrayendo, hay un mayor porcentaje de momentos en los que pierden la vista de la carretera y la transpiración es significativamente mayor que la media.

Como se puede observar para k=2 los grupos obtenidos son fácilmente explicables. Sin embargo a medida que aumentamos el valor de k, los grupos obtenidos se centran más en particularidades de una sola variable y menos en el comportamiento general del grupo. Ahora se aplicarán técnicas de validación de agrupaciones para determinar la cantidad correcta de grupos a partir de medidas estadísticas.

5.1.3 Validación de los agrupamientos

Dado que los agrupamientos son procesos no supervisados de reconocimiento de patrones y cuyos resultados son especialmente sensibles a los parámetros de entrada, es conveniente evaluar el resultado mediante técnicas que proporcionen información sobre cómo se ha formado ese conjunto además de la propia evaluación que realice el analista.

Se diferencian principalmente dos tipos de validación:

- **La validación externa** que utiliza información que no es producto de la propia técnica de agrupamiento utilizada sino que proviene de fuentes externas. Por ejemplo cuando se conoce de antemano a qué clase pertenece cada dato y se realiza una agrupación mediante algún algoritmo, dicho agrupamiento de los datos, con frecuencia será diferente al que indican las clases conocidas de antemano. Se crea entonces una tabla que clasifica si cada objeto ha sido correctamente agrupado en cada grupo según las clases conocidas de antemano:

    1. Verdadero positivo (VP): los puntos que fueron ubicados por el algoritmo en el mismo grupo en que se encontraban según las clases conocidas de antemano.
    2. Falso positivo (FP): aquellos puntos pertenecientes a otros grupos que fueron ubicados por el algoritmo en el grupo dado, que es diferente al que pertenecían de antemano.
    3. Falso negativo (FN): aquellos puntos pertenecientes al grupo dado que fueron ubicados en otro grupo al que no pertenecían inicialmente.
    4. Verdadero negativo (VN): aquellos elementos que fueron ubicados fuera del grupo que comprobamos y que inicialmente no pertenecían a dicho grupo.

    A partir de estas tablas se obtienen métricas como la F-Medida, que se obtiene a partir de los parámetros auxiliares [22]:

    - Precisión: que mide la cantidad de objetos que verdaderamente pertenecen al grupo observado, frente a los que han sido agrupados en dicho grupo. Se calcula:

    $$\text{Precisión} = \frac{VP}{VP + FP}$$

    - Cobertura: que mide la cantidad de objetos que verdaderamente pertenecen a dicho grupo frente a los objetos totales que originalmente pertenecían a dicho grupo. Se calcula:

    $$\text{Cobertura} = \frac{VP}{VP + FN}$$

    - La F-medida puede ser interpretada como la media armónica de los dos parámetros anteriores, aunque utiliza además un parámetro α para regular la preferencia por la medida de la precisión o la cobertura.

$$F\alpha = \frac{1 + \alpha}{\frac{1}{precisión} + \frac{\alpha}{cobertura}}$$

Para $\alpha = 1$ esta medida devolverá la media armónica.

Para $\alpha \in (0:1)$ predominará la precisión.

Para $\alpha > 1$ predominará la cobertura

También se pueden utilizar otras técnicas de validación externa, como la entropía, la pureza o la información mutua, pero solo serán mencionadas, debido a que nuestros conjuntos de datos no están etiquetados y haremos uso de las técnicas de validación externa.

- **La validación interna** mide la eficacia del agrupamiento basada únicamente en la información proporcionada por los datos agrupados por el algoritmo, es decir, proporciona una medida de lo buena que es una estructura de agrupamiento a partir de la información del propio algoritmo y su resultado. La mayoría de estos métodos asignan mejores puntuaciones a los algoritmos que producen grupos con alta similitud entre sus componentes y baja similitud con los otros grupos. Sin embargo, buenas puntuaciones en dichas métricas no equivalen necesariamente a un buen resultado de recuperación de información [22].

  Las métricas de validación interna están basadas usualmente en dos criterios:

  - Cohesión: es una medida de la cercanía entre los miembros de una misma agrupación. Cuanto más cercanos sean los miembros de un mismo grupo (menos distancia existe entre ellos) más cohesionado será el grupo.
  - Separación: es una medida de lo diferentes que son los grupos entre sí. Cuanto más separados estén entre ellos, más diferentes serán los grupos entre sí. Se pueden utilizar diferentes criterios para medir esta magnitud, distancia entre los miembros más cercanos, distancia entre los miembros más distantes o distancia entre centroides.

  A continuación se enumerarán algunas de las técnicas más comunes de validación interna:

  - **Índice de Davies-Bouldin**

    Se obtiene de la fórmula:

$$DB = \frac{1}{k}\sum_{i=1}^{k} \max_{j \neq i}\left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)}\right)$$

    Donde k es el número de grupos, σ_i es la distancia media entre cada punto en el grupo i y su centroide, σ_j es la distancia media entre cada punto en el grupo j y su centroide y $d(c_i, c_j)$ es la distancia entre los centroides de los dos grupos. Cuanto menor sea el coeficiente obtenido, mejor será el algoritmo, bajas distancias intra-grupo (alta semejanza entre elementos de un grupo) y alta

distancia inter-grupo (baja semejanza entre grupos) producen menores índices de Davies-Bouldin [23].

- **Índice de Dunn:**

  Se obtiene de la siguiente fórmula:

$$D = \frac{min_{1 \le i \le j} d(i,j)}{max_{1 \le k \le n} d'(k)}$$

  Donde d(i,j) representa la distancia entre grupos i y j, y d '(k) mide la distancia intra-grupo del grupo k. La distancia inter-grupo d(i,j) puede ser tomada entre cualquier par de elementos de los grupos, así como entre los centroides de los grupos, y la distancia intra-grupo d'(k) puede ser medida como la distancia entre cualquier par de elementos del grupo k. Una mayor puntuación en dicho índice equivale a algoritmos que generan grupos con alta semejanza intra-grupo, y baja semejanza inter-grupo, por tanto, cuanto mayor sea el resultado obtenido, mejor será el algoritmo utilizado [24].

- **Índice de Silueta:**

  El índice de silueta compara la distancia media entre los elementos del mismo grupo con la distancia a elementos de otros grupos. Una puntuación alta indica un objeto bien agrupado, mientras que una puntuación baja puede ser una anomalía. Este índice se utiliza con frecuencia para determinar el número óptimo de grupos [23].

5.1.4 Aplicación de las técnicas de validación

En este proyecto se utilizarán para la validación de los agrupamientos realizados el promedio de distancia dentro del grupo y el índice de Davies-Bouldin. Para ello utilizamos el operador "Cluster Distance Performance" y seleccionaremos ambos métodos de validación de los agrupamientos. Los resultados quedan recogidos en la siguiente tabla:

K	Promedio de distancia dentro de los grupos	Promedio de distancia dentro del grupo 0	Promedio de distancia dentro del grupo 1	Promedio de distancia dentro del grupo 2	Promedio de distancia dentro del grupo 3	índice Davies-Bouldin
2	12,854	10,424	19,853	-	-	2,771
3	11,931	16,167	10,389	18,805	-	2,083
4	11,038	10,319	9,359	18,684	15,434	1,949

Tabla 9 Métricas de validación para los agrupamientos K-means (elaboración propia, 2018)

Como se puede observar para las agrupaciones K-means realizadas, el promedio de la distancia dentro de cada grupo decrece a medida que aumentamos k, esto quiere decir, que los grupos son más compactos, hay más cohesión entre sus elementos. Por tanto desde el punto de vista de este análisis el agrupamiento K-means con k=4 ha generado grupos con miembros más parecidos entre sí y desde este punto de vista es más óptimo que los dos anteriores (k=2 y k=3).

En segundo lugar si se observan los coeficientes del índice de Davies-Bouldin obtenidos se puede apreciar que el coeficiente se reduce a medida que aumenta k, dicho valor indica conjuntamente tanto la separación entre grupos como la compactación de los mismos de manera que cuanto menor es el índice más alejados están los grupos entre sí y más compactos son. Por tanto este coeficiente al igual que la medida anterior indica que el agrupamiento con k=4 agrupa los objetos de mejor modo que los otros dos (k=2 y k=3).

Si se obtienen los valores de los índices para valores de k superiores se puede observar que ambas medidas siguen disminuyendo. Esto es lógico puesto que a medida que añadimos grupos las agrupaciones están formadas cada vez por individuos más similares entre sí. Y por tanto las agrupaciones son consideradas cada vez mejores. Pero esto no es una medida de la cantidad de información extraíble.

En nuestro caso conviene indicar que aunque los objetos agrupados por K-means con k=4 formen grupos más compactos y más distintos entre ellos el resultado obtenido parece más una disección de los datos formando grupos que una clasificación apropiada de las reacciones de los sujetos a las distintas distracciones. Y dado que estas medidas no se relacionan con la relevancia o calidad de las conclusiones obtenidas, es trabajo del analista, utilizando el conocimiento que tenga del problema, decidir cuáles grupos son significativos y cuáles no.

En este caso y debido a las conclusiones obtenidas en cada agrupamiento, se determina que el agrupamiento más significativo es el K-means con dos grupos, en el que existe un grupo que recibe distracciones, sus medidas biométricas aumentan a causa del estrés producido por atender varias tareas a la vez y su atención a la carretera disminuye y un segundo grupo que no recibe distracciones, sus medidas biométricas no aumentan porque no responden a ningún estímulo externo y su atención en la carretera es mayor.

Por lo tanto se obtiene como conclusión, que de modo general la mayoría de distracciones que son orales dentro del mismo coche o bien mediante teléfono móvil, causan una respuesta biológica medible en los conductores como resultado del estrés producido, y una pérdida en la atención observable en el número de veces que apartan los ojos de la carretera. Sin embargo aunque se han diferenciado las distracciones aplicadas en grupos muy distintos, la respuesta medida en las mismas variables biométricas no es suficientemente distinta como para realizar grupos específicos de respuesta a cada estímulo.

Para extraer conclusiones sobre cuáles son las distracciones que mayor perjuicio causan a la atención en la conducción se utilizarán posteriormente los datos (subjetivos) de las encuestas NASA-TLX que los mismos sujetos del estudio han realizado, evaluando la

dificultad de cada una de las pruebas de conducción (en las que cambia la distracción aplicada en cada una, pero no se modifica la dificultad de la actividad de conducción en si misma) pudiendo diferenciar así la apreciación que los sujetos tienen de cada estímulo.

5.1.4 Algoritmo de agrupamiento: X-means

A continuación se realizará otra segmentación de los datos que complemente la validación realizada anteriormente. Para ello utilizaremos el algoritmo X-means el cual implementa mejoras sobre el algoritmo K-means, como la selección automática del número óptimo de centroides. El algoritmo comienza con un número de centroides mínimo y de manera iterativa comprueba si el uso de más centroides es adecuado para los datos. Si un grupo se separa en dos sub-grupos es determinado por el criterio de información bayesiana (BIC), equilibrando el intercambio entre precisión y complejidad de modelo.

Funcionamiento: El algoritmo consta de tres pasos que se repiten hasta que se completa el objetivo. Se eligen un número de grupos mínimo y máximo antes de comenzar, y en la primera ejecución, se comienza con la cantidad de grupos mínimo:

1. Mejorar parámetros: durante este proceso se ejecuta el algoritmo K-means hasta que converge.
2. Mejorar estructura: esta operación averigua si se deberían utilizar nuevos centroides, para ello se dividen los centroides ya existentes en dos grupos que son movidos una distancia proporcional al tamaño de la región que ocupa el grupo original en direcciones opuestas, eligiendo un vector aleatorio para el desplazamiento. Se utiliza entonces el algoritmo K-means de forma local, con k=2 en la región original existente antes de dividir los centroides de forma que los nuevos grupos únicamente se disputan las muestras pertenecientes a la región formada por el grupo antes de dividirse. Por último se evalúa mediante el criterio de información bayesiana, cuál de las dos opciones, si el grupo original o los dos sub-grupos, recoge mejor el comportamiento de los datos en cada una de las regiones y solo se mantiene el modelo que se adapte mejor.
3. Si K> K máx: el algoritmo se detiene y se toma el modelo con más puntuación de la búsqueda.

   Si no, se vuelve al primer paso [25].

A continuación se muestra su implementación en Rapidminer:

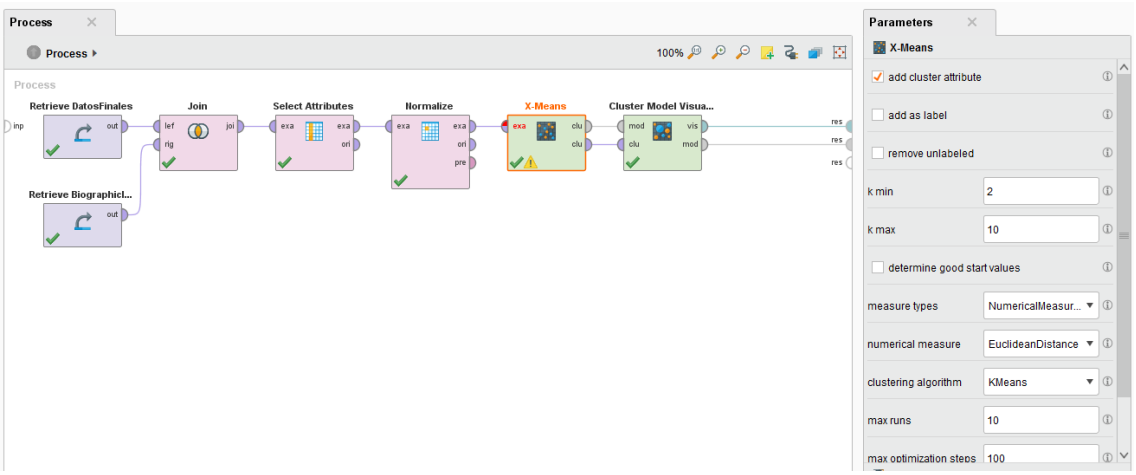


Fig. 41 Implementación del agrupamiento X-means en Rapidminer (elaboración propia, 2018)

Utilizamos la misma configuración que en K-means, pero sustituyendo este operador por X-means, seleccionamos los parámetros k min=2 y k max=20 y por último seleccionamos una métrica distinta, la distancia de Chevychev, como antes se ha mencionado, las agrupaciones realizadas por las técnicas de agrupamiento varían considerablemente en función de la distancia empleada y el número de grupos previamente elegido. Por tanto se intentará encontrar nuevas relaciones utilizando una métrica distinta y un algoritmo que seleccione el número óptimo de grupos de manera automática.

A continuación se muestran los resultados.

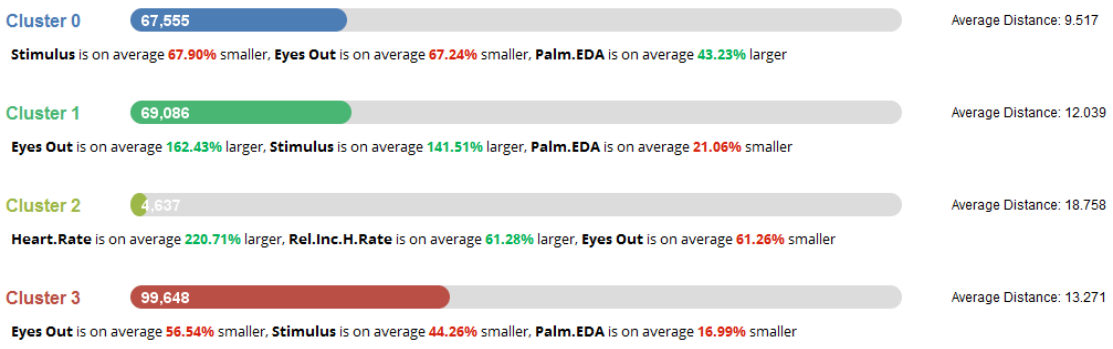


Fig. 42 Vista preliminar del agrupamiento X-means (elaboración propia, 2018)

Attribute	cluster_0	cluster_1	cluster_2	cluster_3
Stimulus	-0.354	0.737	-0.083	-0.231
Palm.EDA	0.337	-0.164	-0.050	-0.132
Heart.Rate	-0.176	-0.140	5.474	-0.036
Breathing.Rate	0.386	0.140	-0.572	-0.346
Perinasal.Perspiration	-0.312	-0.119	-0.284	0.319
Speed	0.124	-0.513	0.201	0.242
Acceleration	0.420	-0.200	-0.030	-0.170
Steering	-0.016	-0.140	0.002	0.105
Lane.Position	-0.261	-0.357	0.016	0.426
Eyes Out	-0.293	0.707	-0.267	-0.246
Rel.Inc.Palm.Eda	0.009	-0.083	-0.410	0.068
Rel.Inc.H.Rate	-0.250	0.001	4.218	-0.018
Rel.Inc.Brething.Rate	0.291	0.159	-0.179	-0.308
Rel.Inc.Perinasal.Pers	-0.331	0.031	-0.076	0.222

Tabla 10 Tabla de centroides del agrupamiento X-means (elaboración propia, 2018)

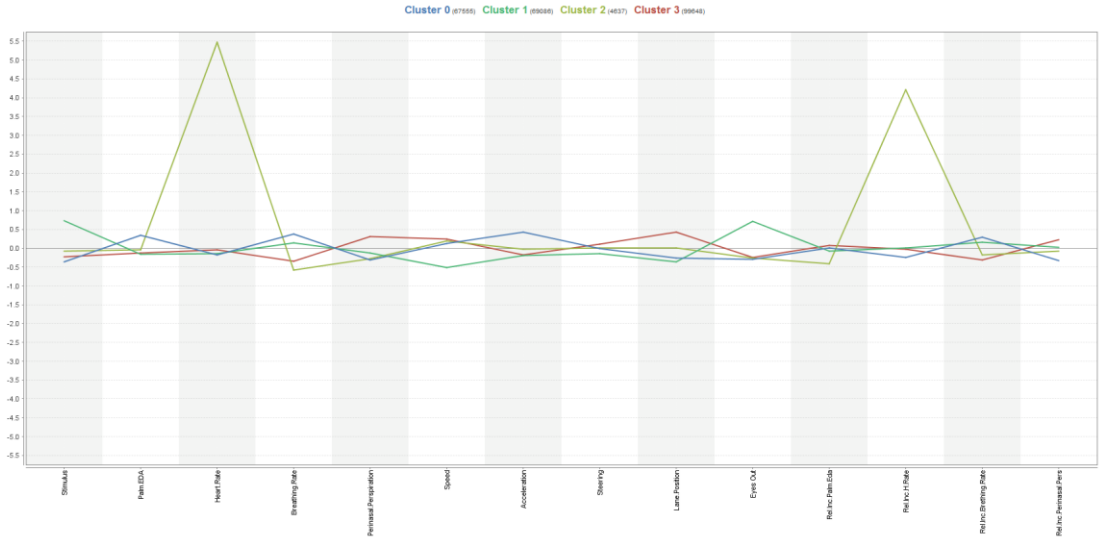


Fig. 43 Gráfica de centroides del agrupamiento X-means (elaboración propia, 2018)

Los resultados observados son notoriamente diferentes a los del agrupamiento K-means k=4 con métrica euclídea.

- Se observa que a los sujetos tanto del grupo 0 como el grupo 3 se les está distrayendo en pocos casos, además el número de veces que pierden visión de la carretera es menor que en los otros grupos, la diferencia entre estos dos grupos radica en que el grupo 0 tiene unos niveles de actividad electrodérmica de las manos altos y el grupo 3 tiene unos niveles bajos. Este atributo no parece estar relacionado con las demás medidas biométricas ni en este ni en los anteriores análisis, y no parece ser un buen indicador de la distracción de los sujetos.
- En el grupo 1 se agrupa a la mayoría de sujetos a los que se está sometiendo a distracciones, los cuales pierden la vista de la carretera un número mayor de veces que el resto de grupos. Esta relación ya se ha observado con anterioridad en los otros análisis.

- En el grupo 2 se agrupan sujetos específicos con un notable aumento del ritmo cardíaco los cuales están atentos a la carretera y la pierden de vista menos veces. Aunque apenas se aplican estímulos a este grupo se puede suponer que el ritmo cardíaco alto es consecuencia del estrés generado en momentos recientes donde sí se les aplicaban estímulos.

Como se puede observar no se está obteniendo información nueva de estos análisis a pesar de utilizar otras distancias. Por tanto se procederá ahora a analizar nueva información obtenida de las encuestas NASA-TLX rellenadas por los sujetos.

5.2 Aplicación de las técnicas de minería de datos a las encuestas NASA-TLX.

5.2.1 Medidas estadísticas

A continuación utilizaremos el fichero que contiene la información psicométrica para determinar qué tipo de distracciones son en opinión de los sujetos las más peligrosas durante la conducción. Recordemos que dicho fichero contiene la valoración subjetiva de cada diferente prueba de conducción, la cual es puntuada de 0 a 20 por cada sujeto en seis diferentes atributos que miden la efectividad en dicha tarea.

En primer lugar se observarán las medidas estadísticas de interés de cada prueba, para obtener una idea general del esfuerzo que ha supuesto para los sujetos realizar correctamente cada prueba.

Para ello utilizaremos el operador "aggregate" de Rapidminer, el cual obtendrá las medidas estadísticas pedidas para cada uno de los grupos formado por todos los sujetos correspondientes a la misma prueba de conducción. Conectamos el operador "aggregate" previamente mencionado y seleccionamos en la lista "aggregation attributes" cada una de las variables evaluadas, demanda mental y física, demanda temporal, rendimiento obtenido, esfuerzo y frustración. Seleccionamos en cada una de ellas la obtención de la media, pues aunque el programa permite obtener múltiples medidas, en este caso conocer "la media de las opiniones" es más relevante, al ser estas ampliamente variables, dado que las capacidades para realizar tareas, varían enormemente de persona a persona.

Por último seleccionamos en "group by attributes" la variable "Driving test" para que aplique la media estadística a los atributos que seleccionamos anteriormente en cada una de las pruebas de conducción.

A continuación se muestra la tabla obtenida.

Row No.	Driving Test	average(Physical Demand)	average(Mental Demand)	average(Temporal Demand)	average(Performance)	average(Effort)	average(Frustration)
1	CD	7.970	14.701	10.239	9.918	14.246	10.082
2	ED	6.209	9.351	7.104	6.313	8.478	5.493
3	FDL	12.844	16.677	14.062	13.922	16.344	14.500
4	FDN	7.286	8.314	7.857	9.029	8.886	8.971
5	MD	11.478	14.224	11.306	11.381	14.836	11.567
6	ND	5.216	6	4.582	4.940	6.507	4.507
7	RD	5.269	7.642	5.075	5.388	8.246	4.955

Tabla 11 Medias de las opiniones de los sujetos frente a cada prueba (elaboración propia, 2018)

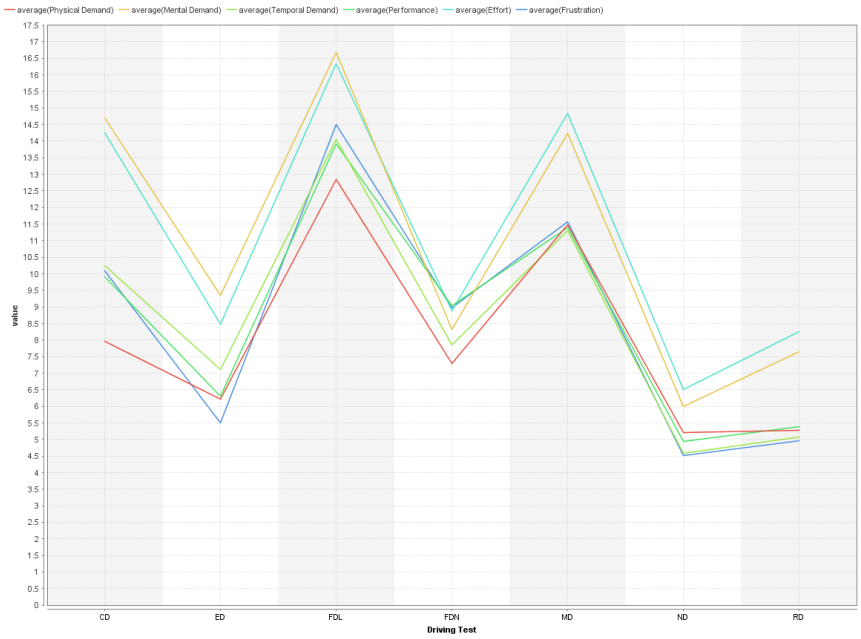


Fig. 44 Gráfica de medias para las opiniones de cada prueba de conducción (elaboración propia, 2018)

Como se puede observar según la gráfica mostrada, hay 3 pruebas que son puntuadas con valores más altos (y por tanto percibidas como más difíciles y frustrantes) y que se diferencian claramente del resto, las cuales son:

- En primer lugar la prueba FDL (conducción con fallo cargada) prueba, que consiste en simular un fallo en el motor en forma de aceleración no intencionada a la vez que se aplica uno de los otros estímulos. Como se puede observar las puntuaciones realizadas por los sujetos a la prueba FDN, la cual aplica solo la aceleración no intencionada y ningún estímulo extra, se puntúa con un nivel mucho más bajo en todos sus atributos. Por tanto se puede inferir que esta prueba es percibida como más difícil por saturar de información a los sujetos y tener que destinar una menor parte de su capacidad total a conducir correctamente.
- En segundo lugar la prueba MD (conducción sensoriomotora) en la cual se distrae a los sujetos mediante mensajes de texto, o llamadas. Esto obliga a los sujetos a apartar en algunos momentos la vista de la carretera para atender los mensajes, por tanto, elimina por algunos instantes el sentido más desarrollado del ser humano (la visión), obligándole a reaccionar más rápidamente ante situaciones inesperadas. Esta es la prueba en la que se utiliza un solo estímulo en la que se obtienen los valores más altos de frustración y esfuerzo requerido. Por tanto se puede deducir que dicha distracción es percibida como la más peligrosa.
- En tercer lugar la prueba CD (conducción cognitiva), es percibida como la que más demanda mental requiere, dicho atributo se encuentra muy ligado al esfuerzo en todas las pruebas, por lo que este también obtiene puntuaciones altas. El resto de atributos obtienen sin embargo puntuaciones mucho más bajas, entre ellos la frustración, lo cual puede ser debido, a que la prueba es

tomada casi como un juego, responder preguntas analíticas y matemáticas mientras se conduce no es una situación normal.

- En valores mucho más bajos se encuentra la última de las pruebas en las que se distrae con un estímulo. ED (conducción emocional), la cual obtiene valores de esfuerzo y frustración solo ligeramente mayores a las pruebas de conducción sin estímulos, posicionándose como una distracción muy leve en la opinión de los sujetos.
- En último lugar se encuentran las pruebas ND (conducción normal) y RD (conducción relajante), las cuales obtienen los valores más bajos de puntuación en la escala NASA-TLX y muy similares entre ambas.

5.2.2 Algoritmo de agrupamiento K-means

Se procederá ahora a realizar de nuevo técnicas de segmentación sobre los datos de las encuestas a fin de obtener más información relevante sobre la impresión de los sujetos sobre los diferentes estímulos estresantes. Se procederá ahora a la implementación del agrupamiento K-means de la misma manera que se hizo con los datos del simulador.

Los resultados se muestran a continuación.

Attribute	cluster_0	cluster_1
Volunteer	33.624	34.423
Driving Test	2.930	4.312
Mental Demand	6.359	15.601
Physical Demand	4.326	11.460
Temporal Demand	4.587	12.246
Performance	5.310	11.495
Effort	6.768	15.331
Frustration	3.958	12.632

Tabla 12 Tabla de centroides para K-means k=2 utilizando los datos de las encuestas (elaboración propia, 2018)

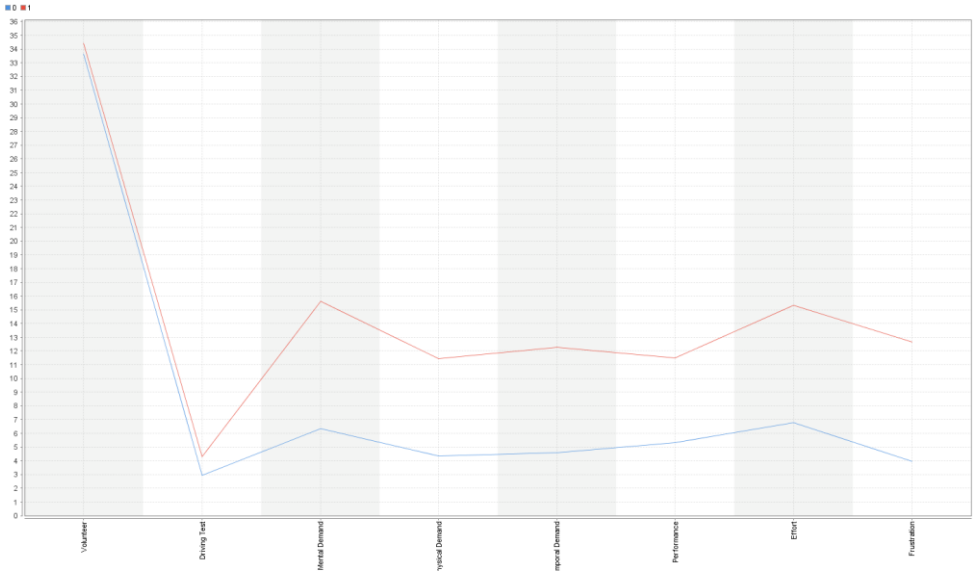


Fig. 45 Gráfica de los centroides para K-means k=2 utilizando los datos de las encuestas (elaboración propia, 2018)

Como se puede observar la clasificación separa dos grupos claramente diferenciados:

El primero es el que obtiene puntuaciones bajas prácticamente en todos los atributos de la escala NASA-TLX. Está compuesto principalmente por las puntuaciones que se da a las pruebas RD (conducción relajante), ND (conducción normal), ED (conducción emocional) y FDN (conducción con fallo normal), también aparecen algunas de las otras pruebas si la puntuación general es baja.

El segundo grupo está formado por las pruebas que obtienen puntuaciones altas en casi todos los atributos de la escala NASA-TLX. A él corresponden principalmente las puntuaciones dadas a las pruebas CD (conducción cognitiva), MD (conducción sensoriomotora) y FDL (conducción con fallo cargada).

Como podemos observar el algoritmo ha clasificado las pruebas en dos tipos, pruebas fáciles y pruebas difíciles según la puntuación de los sujetos. También se puede observar que cuando una prueba es considerada difícil por un sujeto, todos los atributos de la escala NASA-TLX son altos aunque no tengan relación directa entre ellos. Es notable que cuando un sujeto percibe la prueba como difícil y por tanto el esfuerzo necesario es alto el rendimiento que percibe de sí mismo también es superior. Del mismo modo también perciben un rendimiento bajo de sí mismos cuando la prueba requiere poco esfuerzo.

Se han realizado múltiples agrupaciones más utilizando el algoritmo de agrupamiento X-means que como se ha explicado anteriormente determina automáticamente el número óptimo de grupos según la información del conjunto de datos y se han probado diferentes métricas, en concreto la métrica Euclídea, la métrica de Manhattan y la métrica de Chevychev. En todas las pruebas se han obtenido cuatro grupos que dividían las opiniones de los sujetos en cuatro franjas de dificultad. La franja más difícil, contenía siempre los datos de la prueba FDL (conducción con fallo cargada), así como las puntuaciones de otras pruebas cuando estas correspondían a valores muy altos. Los siguientes grupos contenían una mezcla heterogénea del resto de pruebas agrupadas por valores altos, intermedios y bajos. No se ha podido extraer más información útil de estos análisis.

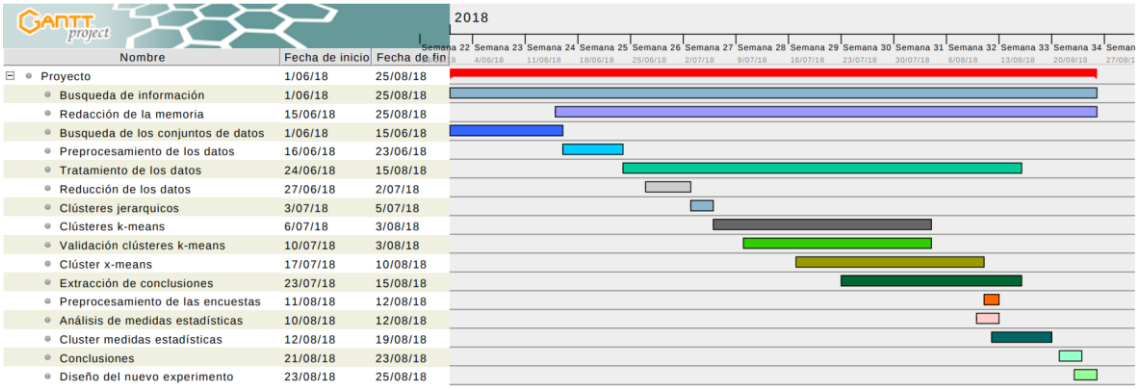
6. Planificación y presupuesto.

En este capítulo se detallarán la planificación inicial del proyecto, que se ha realizado mediante un diagrama de Gantt, así como los costes derivados de la realización de dicho proyecto.

6.1 Planificación

Para gestionar la planificación del proyecto se ha optado por realizar un diagrama de Gantt, dado que el número de tareas a planificar no es muy elevado, y no es necesario un diagrama de PERT que permita visualizar el camino crítico, puesto que las tareas que tienen relaciones de precedencia con las demás son fácilmente identificables, como la obtención de los datos, para su posterior preprocesamiento, y después de este la realización de cualquier intento de análisis.

A continuación se mostrarán las actividades contenidas en el diagrama de Gantt y se comentarán algunos de los objetivos:



Nombre	Fecha de inicio	Fecha de fin
Proyecto	1/06/18	25/08/18
Busqueda de información	1/06/18	25/08/18
Redacción de la memoria	15/06/18	25/08/18
Busqueda de los conjuntos de datos	1/06/18	15/06/18
Preprocesamiento de los datos	16/06/18	23/06/18
Tratamiento de los datos	24/06/18	15/08/18
Reducción de los datos	27/06/18	2/07/18
Clústeres jerarquicos	3/07/18	5/07/18
Clústeres k-means	6/07/18	3/08/18
Validación clústeres k-means	10/07/18	3/08/18
Clúster x-means	17/07/18	10/08/18
Extracción de conclusiones	23/07/18	15/08/18
Preprocesamiento de las encuestas	11/08/18	12/08/18
Análisis de medidas estadísticas	10/08/18	12/08/18
Cluster medidas estadísticas	12/08/18	19/08/18
Conclusiones	21/08/18	23/08/18
Diseño del nuevo experimento	23/08/18	25/08/18

Fig. 46 Diagrama de Gantt del proyecto (elaboración propia, 2018)

Como se puede observar se han destinado dos semanas a la búsqueda de un conjunto de datos apropiado, dado que las variables contenidas en el conjunto tienen que proporcionar información útil en relación a los objetivos fijados.

La redacción de la memoria y la búsqueda de información son actividades que se realizan a lo largo de todo el proyecto por tanto su duración coincide con la duración total del mismo.

Así mismo el procesado o tratamiento de los datos es una actividad que se mantiene casi hasta el final de los últimos análisis realizados, puesto que constantemente se cambia la selección de variables, o se realizan modificaciones a los datos de cara al análisis que vayamos a realizar.

La duración de la realización de los agrupamientos jerárquicos es de solo tres días, porque el tiempo de procesamiento es enorme en relación a los otros tipos de agrupación, el primer intento duró 12 horas de procesamiento hasta que se paró manualmente para modificar el tratamiento de los datos. El segundo intento duró casi 40 horas y no terminó por falta de memoria en el ordenador. Por tanto se desistió de realizar más este tipo de agrupamientos al existir otros algoritmos rápidos y fiables como el algoritmo K-means.

6.2 Presupuesto

A continuación se detallarán los costes de la realización de este trabajo de fin de grado que han sido reducidos al mínimo posible utilizando los recursos gratuitos que se ponen a disposición de estudiantes e investigadores.

Los gastos totales derivados de la realización de un proyecto de minería de datos provienen principalmente de cuatro factores, el coste de la obtención de los datos (si se realiza la toma de los mismos), el coste de la licencia del programa para analizarlos, el coste del hardware empleado y el coste de las horas de trabajo realizadas por los trabajadores.

En principio la obtención de los datos es en sí costosa. En el caso de los datos utilizados en este proyecto se necesitaba disponer de un simulador realista muy costoso, sensores de presión, cámaras, un arnés para tomar el ritmo cardíaco y la respiración, software de reconocimiento de imágenes, así como posiblemente dotar de alguna compensación económica a los voluntarios. Sin embargo algunas organizaciones se dedican a la realización de estudios cuyos objetivos son proporcionar datos mediante medios rigurosos que posteriormente ponen a disposición de investigadores de todo el mundo en páginas dedicadas al almacenamiento y divulgación de dichos conjuntos para que estos obtengan información de ellos.

Para la realización de este trabajo, la obtención propia de datos resultaba imposible por motivos económicos, de modo que se ha ceñido a la búsqueda de conjuntos de datos gratuitos en los cuales se dispusiese de información referente a las causas de accidentes. Dicha búsqueda se ha realizado en sitios web como Kaggle u OSF, encontrando finalmente los datos del estudio: “SIMULATOR STUDY I: A Multimodal Dataset for Various Forms of Distracted Driving” los cuales se pueden utilizar gratuitamente para extraer conclusiones.

En segundo lugar el software de minería de datos utilizado es Rapidminer. Dicho programa ofrece licencias software según las necesidades del cliente. Originalmente los datos utilizados constaban de aproximadamente 300.000 filas de datos. La licencia gratuita solo permite procesar 10.000 filas de datos a la vez, por lo tanto necesitábamos una licencia superior. La licencia pequeña para empresa permite el procesamiento simultáneo de 100.000 filas de datos y tiene un coste de 2,500$ (aproximadamente 2.155€), tampoco podíamos procesar nuestros datos con ella. La licencia media para empresa, permite procesar 1.000.000 de filas de datos simultáneamente pero tiene un coste de 5,000$ (aproximadamente 4.310€), aunque esta licencia nos permitía procesar los datos su coste era demasiado elevado. Por lo que se recurrió a la opción de solicitar una licencia de estudiante, la cual permite procesar datos ilimitados y es gratuita para los estudiantes que se encuentran afiliados a una institución de enseñanza superior.

En cuanto a los equipos utilizados para el tratamiento de datos, se ha utilizado un ordenador de sobremesa, Intel i7 4470k, Nvidia GeForce gtx 770, 16GB ram, 1 TB, del que ya disponía el alumno, comprado hace 4 años. Por tanto el gasto derivado de su uso se obtendrá mediante la siguiente fórmula de amortización:

Cálculo de amortización del equipo informático:

$$\frac{A}{B} x C x D$$

Siendo:

A= nº de meses desde la fecha de facturación en que el equipo es utilizado (4 meses)
B= periodo de depreciación (96 meses, equivalente a 8 años)
C= coste del equipo sin IVA (1.054,98€)
D= % del uso que se dedica al proyecto (aproximadamente 90%)

$$Total = \frac{4}{96} x\ 1.054{,}98\ x\ 0{,}9 = 39{,}56€$$

Por último falta contabilizar el coste de las jornadas de trabajo realizadas por el alumno, que corresponderían al puesto de Analista Programador, así como las jornadas de trabajo del tutor que corresponderían al puesto de Ingeniero/Consultor Senior. Usando un promedio de las tarifas de varias consultoras para los perfiles de Analista Programador e Ingeniero/Consultor Senior durante el año 2018, el cálculo del precio por jornada y por perfil resulta de 186€/jornada de Ingeniero Senior y 70€/jornada de Analista Programador. Con estos datos calcularemos el coste de las horas de trabajo realizadas:

$$110 \text{ jornadas de trabajo de analista programador x } 70€ = 7700€$$

$$5 \text{ jornadas de trabajo de Ingeniero Senior x } 186€ = 930€$$

A continuación se resumen los costes de realización del proyecto en el presupuesto:

Presupuesto	
Obtención de los datos.	0 €
Amortización del Hardware utilizado	39,56 €
Licencia Rapidminer Estudiante	0 €
Jornadas de trabajo del alumno	7700 €
Jornadas de trabajo del tutor	930 €
Total	**8669,56 €**

Tabla 13 Presupuesto (elaboración propia, 2018)

7. Conclusiones y trabajos futuros

En este apartado se expondrá un sumario de las conclusiones obtenidas durante la realización de los análisis, la medida de la consecución de los objetivos, así como recomendaciones para la realización de estudios posteriores basados en la experiencia obtenida en la realización de este estudio.

7.1 Conclusiones

En el caso de este proyecto, se ha obtenido las siguientes conclusiones de los conjuntos analizados:

- Las agrupaciones realizadas en los datos del simulador han mostrado que la aplicación de los estímulos causa una respuesta medible en las medidas biométricas de los voluntarios.
- Al ser los diferentes estímulos aplicados de una manera similar (o bien oralmente, o bien mediante teléfono móvil) las respuestas de los voluntarios no se diferenciaban suficientemente como para que las agrupaciones naturales pudieran crear grupos basados en cada estímulo.
- El principal factor de riesgo detectado es la pérdida de visión de la carretera, cuya ocurrencia era muy superior cuando se aplicaban estímulos.
- Según la percepción de los voluntarios, las pruebas más difíciles de realizar (de un solo estímulo) eran la prueba sensoriomotora, en la cual tenían que leer y escribir mensajes de texto, o hablar por teléfono, y la prueba cognitiva en que tenían que responder a preguntas analíticas y matemáticas. En el caso de la prueba sensoriomotora el uso del teléfono tiene como consecuencia tener que apartar la visión de la carretera, por lo que esa pérdida de información visual momentánea obliga a reaccionar más rápidamente a cualquier cambio inesperado. De la misma manera contestar las preguntas analíticas puede ocupar parte de la capacidad analítica de los sujetos que de otro modo estaría destinada a la actividad de conducción. Ambos estímulos pueden ser considerados factores de riesgo y requerir estudios adicionales.
- La prueba de conducción emocional era percibida prácticamente con los mismos valores que la conducción normal o relajante, de modo que esta distracción es percibida como muy poco peligrosa.
- Se han observado diferencias significativas entre las medidas biométricas obtenidas por los sujetos según su edad, pero ninguna de las agrupaciones ha mostrado diferencias entre los distintos sexos.

Aunque originalmente se seleccionó un set de datos con muchísimas muestras (casi 300.000 líneas de datos) y otros conjuntos que aportaban información complementaria, las técnicas de aprendizaje no supervisado (agrupamiento) que hemos aplicado a este set de datos sin etiquetas no han obtenido la cantidad de información que se esperaba. La cantidad de datos no es proporcional a la cantidad de información que se va a obtener.

Las técnicas de agrupamiento suelen tener como objetivo dividir o segmentar los conjuntos de datos en agrupaciones homogéneas a las que sea más fácil aplicar un modelo

y la cantidad de información que aportan es menor que otras técnicas. Además es difícil definir cuando el resultado de un agrupamiento es aceptable porque aunque los métodos de validación de los agrupamientos utilicen métodos estadísticos para definir la calidad de la agrupación, esto no garantiza la calidad de la información obtenida.

En este caso se han intentado encontrar más conjuntos de datos de los que se pudiera obtener información relativa a las causas de distracción al volante para completar la información obtenida, pero no ha sido posible encontrar otro set de datos que aportara nueva información relativa a la búsqueda de factores de riesgo en la conducción.

7.2 Trabajos futuros

Como se explicó al inicio de este proyecto, los accidentes de tráfico causan enormes pérdidas tanto para los accidentados, como para las economías de los lugares donde estos ocurren, por tanto es de interés para los gobiernos reducir las cifras de accidentes invirtiendo en investigación y prevención. Por ello desde aquí se propone un plan de toma de datos relativo a las causas de accidentes, destinado a ser financiado por países de economías poderosas u organizaciones de las que forman parte múltiples países como la Unión Europea.

El proyecto consistiría en una reducción de precio en vehículos parcialmente financiados por la organización investigadora que contasen con un sistema electrónico integrado de toma de datos que se guardarían de manera local en una caja negra en el propio vehículo. Dichos datos no podrían ser recogidos ni analizados salvo en caso de accidente, caso en el cual los datos solo podrían ser tratados con fines de investigación y no con fines jurídicos o legales, protegiendo así la privacidad de los usuarios de este sistema.

La información almacenada solo correspondería a las últimas 6 horas de conducción y los datos se irían borrando a medida que llegasen datos nuevos. El estudio estaría diseñado para ser lo menos intrusivo posible, por tanto se utilizarían muchas técnicas de reconocimiento de imágenes para reducir al mínimo la interacción directa con el sistema, pero siendo todos los datos recogidos de carácter real, y de manera más importante, el conjunto de datos obtendría muestras de las conductas con las cuales se causan accidentes reales. Para ello el número de vehículos disponibles debe ser suficientemente grande. Y aunque no se obtengan datos inmediatamente debido a que cada año solo se sacarían mediante este plan un número limitado de vehículos, con el paso de los años se irían acumulando nuevos datos de los vehículos accidentados y progresivamente se iría formando un conjunto de datos con multitud de información para la prevención de accidentes.

Variables a registrar:

- En primer lugar los vehículos estarían dotados de una centralita que recibiese los valores de los sensores que medirán los datos relativos a la conducción, como en el caso de los datos del estudio que hemos usado. Estas variables serían velocidad, aceleración, freno y la posición respecto a la línea mediante una cámara en el lateral derecho del vehículo.

- En segundo lugar se obtendrían mediante reconocimiento de imágenes posteriormente al accidente, el ritmo cardíaco, las variables posición de los ojos y ritmo respiratorio, así como el grado de apertura de los ojos, que quedarían grabadas en video durante la conducción con una cámara interna, en un plano suficientemente amplio para ver la actividad de los acompañantes. Adicionalmente se utilizaría un sistema "Drive alert" que enviaría señales mediante bluetooth al coche cuando el sujeto inclinase la cabeza hacia delante en un ángulo que desfavoreciese la visión.
- Por último la cámara contaría además con un sensor de infrarrojos que permitiría medir la temperatura del sujeto.

Mediante estas variables se pretende de manera no intrusiva recoger datos relacionados con los principales factores que influencian los accidentes (según la DGT), según la cual la conducción distraída es uno de los factores de más peso en los accidentes, entre dichas distracciones, la más peligrosa es la distracción visual de la que se ha hecho uso en este proyecto, pero grabar la actividad del sujeto dentro del coche permitiría detectar distracciones auditivas si está hablando con otro sujeto, o biomecánicas (en las que el sujeto esté utilizando el móvil o simplemente ajustando la radio). Y determinar bajo qué condiciones dichas distracciones causan un accidente.

De la misma manera la apertura de los ojos, así como la inclinación de la cabeza serían claros indicadores de somnolencia o fatiga, que es el tercer factor más influyente en accidentes en nuestro país.

Con todo esto se pretende almacenar datos de la actividad de los conductores y de los factores de su entorno cuando realmente ocurre un accidente, de manera que se pueda estimar de manera más precisa en qué medida afecta cada uno de los factores observados a dicho accidente.

Bibliografía

[1] “Las 10 principales causas de defunción”, Organización mundial de la salud (OMS), 2018. Disponible en: http://www.who.int/mediacentre/factsheets/fs310/es/index1.html

[2] “Accidentes con víctimas, fallecidos 30 días, heridos graves y leves”. Series históricas, DGT, 2016.

[3] Josefa Valcárcel, “Las principales cifras de siniestralidad vial, España, 2016” Dirección General de Tráfico, España, Informe técnico NIPO: 128-15-069-X, 2017.

[4] “Balance de seguridad vial”, Dirección General de Tráfico, España, 2017.

[5] J. Han, M. Kamber, J. Pei. *Data Mining Concepts and Techniques.* Tercera edición. Waltham: Elsevier, 2012

[6] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer and R. Wirth, “CRISP-DM 1.0 Step-by.step data mining guide”, CRISP-DM consortium, 1999

[7] C. Pérez y D. Santín. *“Minería de datos: técnicas y herramientas”.* Madrid: Ediciones Paraninfo, S.A. (2007).

[8] Pradnya P. Sondwale, “*Overview of Predictive and Descriptive Data Mining Techniques”* Research Paper, Conference Held at IETE Amravati Center, Maharashtra, India, 2015 Disponible en: http://ijarcsse.com/Before_August_2017/docs/papers/Special_Issue/ITSD2015/58.pdf

[9] Richard Lowry “One-Way Analysis of Covariance for Independent Samples” 1999-2006. Chapter 17.

[10] M. Flak, F. Marohn, R. Michel, D. Hofmann, M. macke, C. Spachmann, S. Englert, “A First Course on Time Series Analysis” GNU Free Documentation License, 2012.

[11] John H. Holland, “Genetic Algorithms”, Scientific American Vol. 267, No. 1 (JULY 1992), pp. 66-73

[12] T. Menzies, Y. Hu. Data Mining For Busy People. IEEE Computer, 2003

[13] Brian Everitt, Cluster analysis. Chichester, West Sussex, U.K: Wiley. (2011).

[14] Malcolm Dcosta. “SIMULATOR STUDY I: A Multimodal Dataset for Various Forms of Distracted Driving.”. Open Science Framework, 2017. Disponible en: https://osf.io/c42cn/

[15] S. Taamneh, P. Tsiamyrtzis, M. Dcosta, P. Buddharaju, A. Khatri, M. Manser, T. Ferris, R. Wunderlich & I. Pavlidis, “A multimodal dataset for various forms of distracted driving”. Scientific Data volume 4, Article number: 170110, 2017. Disponible en: https://www.nature.com/articles/sdata2017110#cite2

[16] I. Pavlidis, M. Dcosta, S. Taamneh, M. Manser, T. Ferris, R. Wunderlich, E. Akleman & P. Tsiamyrtzis, “Dissecting Driver Behaviors Under Cognitive, Emotional, Sensorimotor, and Mixed Stressors” Scientific Reports volume 6, Article number: 25651, 2016. Disponible en: https://www.nature.com/articles/srep25651

[17] “NASA TASK LOAD INDEX (TLX) Paper and pencil Package” NASA Ames research center. California. Disponible en: https://humansystems.arc.nasa.gov/groups/TLX/downloads/TLX_pappen_manual.pdf

[18] J. A. Hartigan and M. A. Wong “Algorithm AS 136: A K-means Clustering Algorithm” Royal Statistical Society. Series C. 1979.

[19] Aurea Grané, Análisis de Componentes Principales, Departamento de Estadística, Universidad Carlos III de Madrid. Disponible en: http://halweb.uc3m.es/esp/Personal/personas/agrane/ficheros_docencia/MULTIVARIANT/slides_comp_reducido.pdf

[20] Richard J. Roiger “Data Mining, A Tutorial-Based Primer”, Segunda edición, Chapman & Hall, 2017.

[21] John A. Hartigan, “Clustering Algotithms”, primera edición. John Wiley & Sons. 1975.

[22] C. D. Manning, P. Raghavan, H. Schütze, “An introduction to information retrieval”, Cambridge UP, Cambridge, England. 2009. Disponible en: https://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf

[23] Bernard Desgraupes “Clustering índices” University Paris Ouest, 2017. Disponible en: https://cran.r-project.org/web/packages/clusterCrit/vignettes/clusterCrit.pdf

[24] J. Dunn, “Well separated clusters and optimal fuzzy partitions”. Journal of Cybernetics, (1974).

[25] D. Pelleg y A. Moore, “X-means: Extending K-means with Efficient Estimation of the number of Clusters” School of Computer Science, Carnegie Mellon University, Pittsburgh. Disponible en: http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=08BA4B033265864A6AD78F51F63E9D3C?doi=10.1.1.19.3377&rep=rep1&type=pdf

[26] Rapidminer Studio Manual. 2014

[27] Rapidminer Training Videos. Disponible en: https://Rapidminer.com/training/videos/

[28] Dr. Matthew A. North, “*Data Mining for the masses*” primera edición. Global Text Project, 2012.

[29] Rapidminer Documentation v 9.0 , K-means, Disponible en: https://docs.rapidminer.com/latest/studio/operators/modeling/segmentation/k_means.html

[30] Rapidminer Documentation v9.0, Normalize. Disponible en: https://docs.rapidminer.com/latest/studio/operators/cleansing/normalization/normalize.html

[31] Eibe Frank, Mark A. Hall, and Ian H. Witten (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.

[32] SAS® Enterprise Miner™, Fact sheet. Disponible en: https://www.sas.com/content/dam/SAS/en_us/doc/factsheet/sas-enterprise-miner-101369.pdf

[33] Orange Documentation v2.7.8. Disponible en: https://docs.orange.biolab.si/2/#

[34] Knime, Documentation. Disponible en: https://www.knime.com/documentation

[35] Jahir A. Gutiérrez O., y B. Molina, "Identificación de técnicas de minería de datos para apoyar la toma de decisiones en la solución de problemas empresariales". Revista Ontare. (2016).

[36] GA Meiring, HC Myburgh, "A Review of Intelligent Driving Style Analysis Systems and Related Artificial Intelligence Algorithms" Department of Electrical, Electronic and Computer Engineering, University of Pretoria, Hatfield, Pretoria, 2015, Disponible en: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4721742/

ANEXO A: Extended Abstract

Introduction

According to the WHO (World Health Organization), traffic accidents caused 18.3 deaths per 100,000 inhabitants worldwide. This figure is much more pronounced in low-income countries than in high-income countries [1]

In Spain according to the historical series of accidents with deceased victims, serious and minor injuries of the DGT (General Directorate of Traffic) between 80,000 and 100,000 annual traffic accidents have been happening for more than two decades [2].

Road accidents have the terrible consequences of the loss of human life or quality of life, which are the most important, but they also have a series of associated costs that burden the economies.

These costs can be broken down into three categories according to a study framed by the COST 313 action executed by the European Commission in the early nineties to review the way in which European countries estimated the costs of accidents and make recommendations on how they should be quantified [3]:

- The direct economic costs: medical costs, costs of repair or replacement of damaged vehicles and administrative costs.
- Indirect costs: the value of productive capacity lost as a result of premature death, or of permanent or temporary disability caused by the loss.
- The value of the lost quality of life: the value of the loss of health or loss of enjoyment of the life of the victim, as well as the pain, affliction and suffering of the victim and their relatives.

In Spain in 2011 the DGT in collaboration with the University of Murcia, estimated the costs associated with traffic accidents with victims using the method of payment arrangement. According to this estimation, a decease would cost 1.4 million €, accounting for direct and indirect costs, as well as the value of a statistical life. In the same way, the costs associated with a hospitalized injured person were estimated at 219,000€ and those of a non-hospitalized injured person at 6,100€. These valuations were updated in 2016, taking as a reference the nominal variation of the gross domestic product per capita, resulting in a total estimated costs of accidents in 2016 of 10,275 million euros [3].

According to the DGT data again, the main contributing factors in serious accidents and their percentage of occurrence as the main cause of the accident are listed below [4]:

Distracted driving	----	32%
Inadequate speed	----	26%
Fatigue or drowsiness	----	12%
Alcohol	----	12%
Drugs	----	11%

It is noticeable that driving distractions cause arround a third of the total serious accidents. The purpose of this study will be to determine the main distractions that influence the performance of the drivers, thus increasing the chances of an accident. To achieve this

objetive we will use data mining techniques in different datasets that will provide usefull information about the distractions used, as well as the response of the subjects both physical and mental (their perception of the distraction).

Data overview

The data for these analyses have been provided by Malcolm Dcosta's study "SIMULATOR STUDY I: A Multimodal Dataset for Various Forms of Distracted Driving." [14]. In this study, 68 volunteers drive on a highway under the influence of different stressors, meanwhile, the data of the driving simulator (brake force, speed, acceleration, etc) is recorded, as well as biometric information (heart rate, breathing rate, eyes position, perinasal transpiration, palm electrodermic activity, etc.). After that, they evaluate the subjective work-load of each driving test using the NASA-TLX scale.

The attributes of the data sets will be explained next:

Biographic information data sets contains the next attributes for each volunteer:

- Sex: Man/woman
- Age: numeric value
- Age Group: Young/old

Simulator data sets contains the driving simulator data, the biometric data, and the information related to the driving test and the stressor applied, the attributes are the following:

- Time: time in seconds to synchronize the other attribute's values.
- Drive: It indicates what driving test is in course. It takes a value in $\mathbb{N}$ =[1,8]
    - 1 = Baseline
    - 2 = Practice Drive
    - 3 = Relaxing Drive
    - [4,7] (random order for each volunteer)
      Normal drive: no stimulus applied.
      Cognitive drive: analytical and mathematical questions.
      Emotional drive: emotional questions.
      Sensorimotor drive: mobile phone text and calls.
    - 8 = Failed drive: unintended acceleration of the vehicle. Can be further loaded with a stimulus or not, being Failed Drive Loaded (FDL), or Failed Drive Normal (FDN).
- Stimulus: A variable to indicate if a stressor is being used during the test in that moment. It takes a value in $\mathbb{N}$ =[0,5]
    - 0 = No stimulus
    - 1 = Analytical questions
    - 2 = Mathematical questions
    - 3 = Emotional questions
    - 4 = Text messages
    - 5 = Text messages and calls.
- Failure: to show if the unintended acceleration is being applied, 0 for normal working of the machine and 6 for unintented acceleration.

- Speed: Speed of the car in the simulator.
- Acceleration: Acceleration of the car in the simulator.
- Brake: Force applied to the brake in N.
- Steering: Steering wheel angle.
- Lane Position: Distance to the right line, positive values are positions of the car at the left of the line. Negative values are positions of the car at the right of the line.
- LaneOffset: Position of the car from the center of the lane.
- Distance: Distance travelled by the car.
- Palm EDA: Electrodermal activity in the palms of the volunteer, an attribute to measure the stress.
- Perinasal transpiration: Transpiration of the perinasal region obtained by image recognition.
- Heart Rate: Value in bpm.
- Breathing Rate: Value in bpm.
- Gaze.X.Pos and Gaze.Y.Pos: Variables to track eye position in coordinates X and Y.

And the psychometric data set contains the subjective perception of each volunteer related to the following attributes for each driving test (NASA-TLX scale [17]):

- Mental demand: How mentally demanding was the task?
- Physical demand: How physically demanding was the task?
- Temporal demand: How hurried or rushed was the pace of the task?
- Performance: How succesful were you in accomplishing what you were asked to do?
- Effort: How hard did you have to work to accomplish what you were asked to do?
- Frustration: How insecure, discouraged, irritated, stressed and annoyed were you?

Each attribute has a numeric value in the range $\mathbb{N} = [1, 20]$.

Variables added for their utility for our study in the simulator data sets:

- EyesOut: A variable used to know if the subjet is not making visual contact with the screen in that moment. It is obtained from Gaze.X.Pos and Gaze.Y.Pos, because the eye tracking system is active all the time, and when it doesn't return a value for the positions it's because it can't locate the eyes of the volunteer, either because he is not looking at the screen of simulator or because he has his eyes closed.
- DistractionOn: A variable to know when an stimulus is being used but not caring about which one it is.
- Rel.Inc.Heart.Rate, Rel.Inc.Breathing.Rate, Rel.Inc.Perinasal.Perspiration and Rel.Inc.Palm.EDA: These variables measure the relative increase or decrease with respect to the average of each subject in a state of relaxation, heart rate, respiration, perinasal perspiration and electrodermal activity of the hands respectively.

Data preparation

The first step to do with our data is cleaning it from all kind of noise, incorrect values, and missing values that can make our results much less accurate. This is an iterative proccess that is done multiple times throughout the study, but we will present a quick summary here.

Our data shows some problems in the statistics, mainly incorrect values (out of the sensor's measuring range or senseless) and a lot of missing values.

So the two principal solutions implemented in this project to correct that problems are:

- **Correct values Filter**: This filter is designed to discard the samples of the dataset that has one attribute with senseless values or out of the sensor's measuring range, or more than one attribute with missing values.
  Its implementation in RapidMiner consists in a succession of "filter examples" operators. The first filter keeps the values of all the variables at the same time, that are correct and don't have missing values. This discards all the simples with missing values.
  The next filter will use the discarded samples of the first filter as input, and will keep the samples that have one of the attributes filtered previously missing, only if the other attributes are in the sensor's measuring range.
  With this second filter the samples with the selected attribute missing are recovered, but this way samples with two or more missing attributes or with incorrect values in them aren't recovered.
  Filters that use as input the previous filter's discarded values keep being added, that will recover the samples of every attribute filtered in the first filter, only if this one is missing and the other values are correct.
  Then all the filtered values are joined in a new set of correct values with the operator "append". The new dataset contains attributes with correct values and a máximum of 1 missing attribute.

- **Missing values handling:** As keeping the maximum variance in our data is recommended, it isn't a good approach to simply remove more samples if they have one missing attribute, and at the same time clustering algorithms can't be used if there are missing values.
  So the best approach is to fill the missing values with the mean of that attribute.
  There must only be taken into consideration that the values of the biometric attributes vary greatly from one volunteer to another. So in our implementation is chosen to fill that gaps with the mean of the values of that attribute, for every volunteer separately.
  Then the only atribute's values missing will be the ones that didn't have any sample with value for a entire volunteer's attribute. As it's prefered to keep the information of all other attributes we fill that gaps with the general mean of all dataset's samples of such attribute.

Data analyses

This section is divided in two parts, the first one consists in the analyses made with the driving simulator data, that are PCA, K-means, and X-means, the second one consists in the analyses made with the psychometric data, that are summarization and K-means.

Simulator data analyses

PCA: As a first step a Principal Component Analysis is made in an attemp to reduce the dimension of our data set. This is a technique that seeks the projection of the data in terms of least squares, according to which the data are represented as best as possible. The data set is described in terms of new uncorrelated variables or "components". This type of analysis makes sense if there are high correlations between the variables, since this is indicative of redundant information, and therefore the whole could be reduced to a few factors that would explain most of the total variability of the data set.

A study of the correlation between variables is made with a correlation matrix, and it shows low correlations between all the attributes of our data set, with a correlation between 0 and 0.4 in most of our attributes, It is considered that there is no correlation between the variables, and although it shows a small interaction between attributes, it is not statistically relevant. That is a clear sign that our data set won't reduce its dimensión a lot if the analysis is applied.

Attribut...	Drive	Stimulus	Failure	Palm.EDA	Heart.R...	Breathi...	Perinas...	Speed	Acceler...	Brake	Steering	LaneOff...	Lane.Po...	Eyes Out	Age
Drive	1	0.464	0.091	-0.070	-0.001	-0.058	0.090	0.024	-0.046	0.017	-0.000	0.098	-0.200	0.104	0.019
Stimulus	0.464	1	0.058	-0.009	0.027	0.103	0.099	0.025	-0.069	0.001	0.000	0.030	-0.203	0.197	-0.006
Failure	0.091	0.058	1	-0.004	0.011	-0.004	0.034	-0.066	-0.060	0.163	-0.002	-0.013	0.007	0.015	-0.002
Palm.EDA	-0.070	-0.009	-0.004	1	-0.082	0.080	-0.048	-0.052	0.052	-0.024	0.006	0.055	-0.010	0.057	0.252
Heart.Rate	-0.001	0.027	0.011	-0.082	1	-0.038	0.075	0.029	-0.004	-0.017	0.002	-0.026	0.004	-0.041	-0.280
Breathin...	-0.058	0.103	-0.004	0.080	-0.038	1	0.028	-0.019	0.002	-0.011	0.001	-0.017	0.071	0.115	-0.134
Perinasa...	0.090	0.099	0.034	-0.048	0.075	0.028	1	-0.032	-0.028	0.045	0.003	0.007	-0.031	0.016	0.163
Speed	0.024	0.025	-0.066	-0.052	0.029	-0.019	-0.032	1	0.097	-0.329	-0.006	0.016	0.065	-0.037	-0.056
Accelerat...	-0.046	-0.069	-0.060	0.052	-0.004	0.002	-0.028	0.097	1	-0.362	0.010	0.006	0.019	0.034	-0.010
Brake	0.017	0.001	0.163	-0.024	-0.017	-0.011	0.045	-0.329	-0.362	1	-0.003	-0.010	-0.014	-0.036	0.103
Steering	-0.000	0.000	-0.002	0.006	0.002	0.001	0.003	-0.006	0.010	-0.003	1	-0.122	0.089	-0.000	0.002
LaneOffs...	0.098	0.030	-0.013	0.055	-0.026	-0.017	0.007	0.016	0.006	-0.010	-0.122	1	-0.070	0.014	0.070
Lane.Po...	-0.200	-0.203	0.007	-0.010	0.004	0.071	-0.031	0.065	0.019	-0.014	0.089	-0.070	1	-0.043	-0.053
Eyes Out	0.104	0.197	0.015	0.057	-0.041	0.115	0.016	-0.037	0.034	-0.036	-0.000	0.014	-0.043	1	0.004
Age	0.019	-0.006	-0.002	0.252	-0.280	-0.134	0.163	-0.056	-0.010	0.103	0.002	0.070	-0.053	0.004	1

Fig. 24 Correlation matrix (own elaboration, 2018)

For learning purposes we end the analysis to see how many new uncorrelated attributes or "components". As we previously assumed by the results of the correlation matrix, of the 15 main attributes that our set has, we would need 14 to obtain a total variance of 96.8%. So it would only reduce one attribute if we use this new components, because of that we won't use this new set attributes.

K-means clustering algorithm

In the search of natural groups that can reveal relation between variables, and groups of behaviours or reactions as response from a stressor we use the algorithm K-means. That will make a preselected amount of groups. In the first attemps we use all the variables availables and we find that some of them are obscuring the results.

The inclusion of the brake (force applied to the brake) to the analysis makes always some groups based on that attribute, because the difference between the values when the brake

is not being used (0 N), and when it's being pressed (hundreds of N) is too high. So the metrics of the algorithm K-means find a great difference between the samples where the brake is pressed when is not.

A similar thing happens with the age attribute, there are two groups of people in similar quantities, one is the young group formed by people in their twenties and and old group formed by people older than sixty. There is a relation between the biometric measures and the age of the volunteers, as we can find in that analyses, but the main objetive of this study is to find how distractions affect drivers capacity in a perceptible way. So we remove both age and brake attributes from this analysis in order to find the relations we search for.

After choosing the correct variables we start with K-means with two groups.

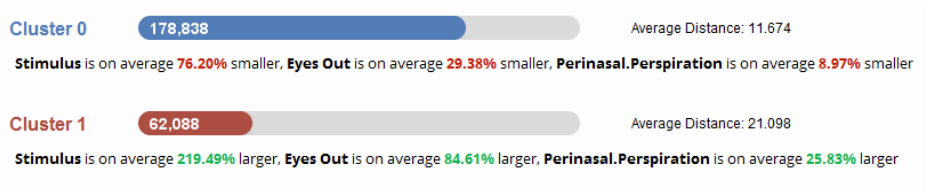


Fig. 47 Overview of k-means, k=2 (own elaboration, 2018)

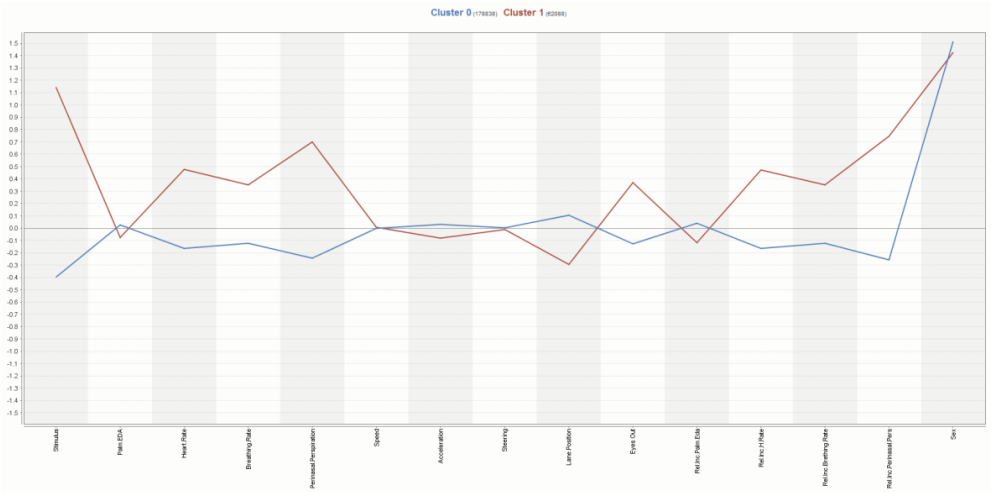


Fig. 36 Centroid chart of k-means, k=2 (own Elaboration, 2018)

As expected, in this case a more relevant role of applied stimulus and biometric magnitudes is obtained. As it can be observed the algorithm has differentiated the two groups mainly by the application of distractions.

Cluster 0 is made up of individuals who are not being distracted, who lose less sight of the road and their biometric measurements are much lower than those of the other group because they do not manifest any stress or anxiety.

Cluster 1, however, corresponds to those samples of the individuals to whom a distraction is being applied and, as a result, they lose more attention from the road and their biometric measurements are higher due to stress.

We repeat the aplication of the algorithm with three groups.

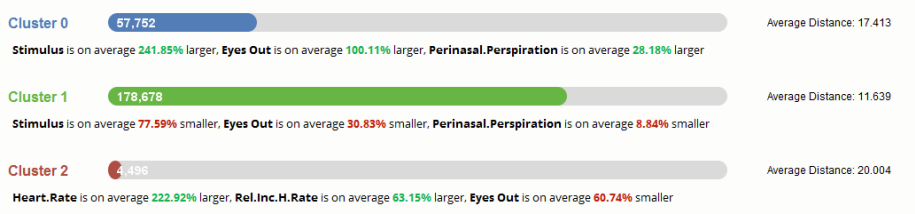


Fig. 37 Overview of k-means, k=3 (own elaboration, 2018)

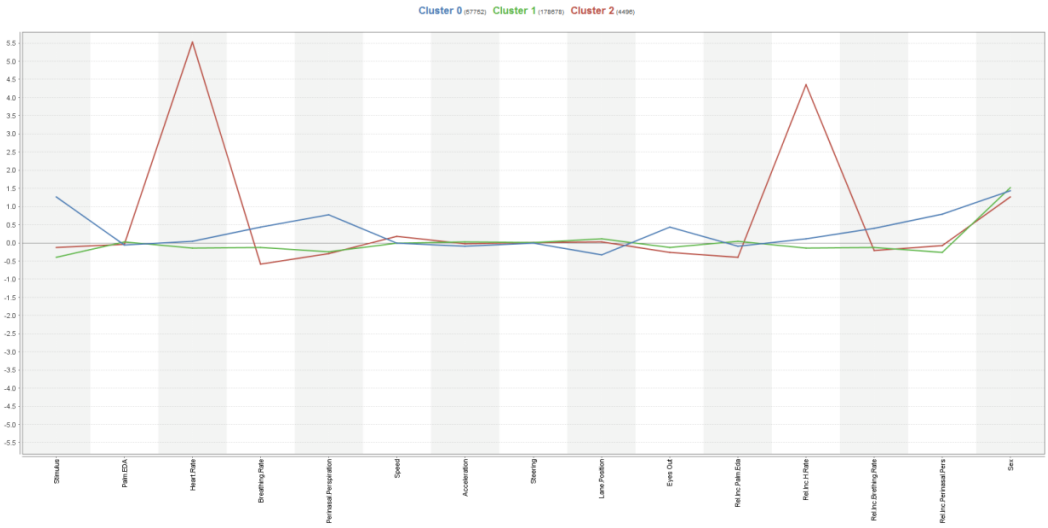


Fig. 38 Centroid chart of k-means, k=3 (own Elaboration, 2018)

As we can see, the groups are similar to the previous clustering, but approximately 4000 samples from group 1 of that analysis have formed a new group (cluster 2).

Cluster 0, corresponds to those samples of the individuals to whom a distraction is being applied and, as a result, they lose more attention from the road and their biometric measurements are higher due to stress.

Cluster 1 is made up of individuals who are not being distracted, who lose less sight of the road and their biometric measurements are much lower.

Cluster 2 corresponds to those samples of individuals who have high heart rate, and this heart rate is increased respect their mean in rest state, indicating they are stressed. This group is however focusing more on the screen.

Once again we repeat the algorithm with four groups.

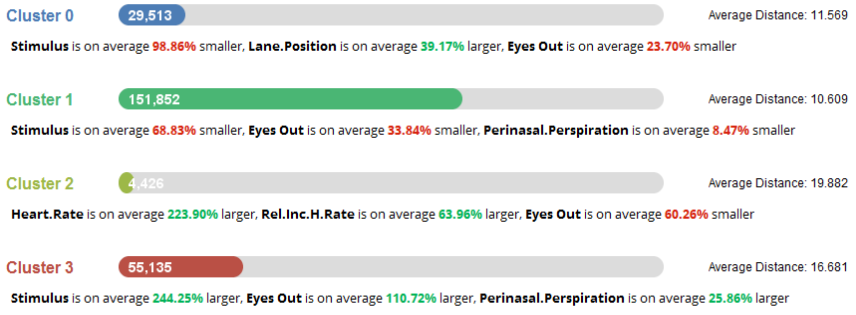


Fig. 39 Overview of k-means, k=4 (own elaboration, 2018)

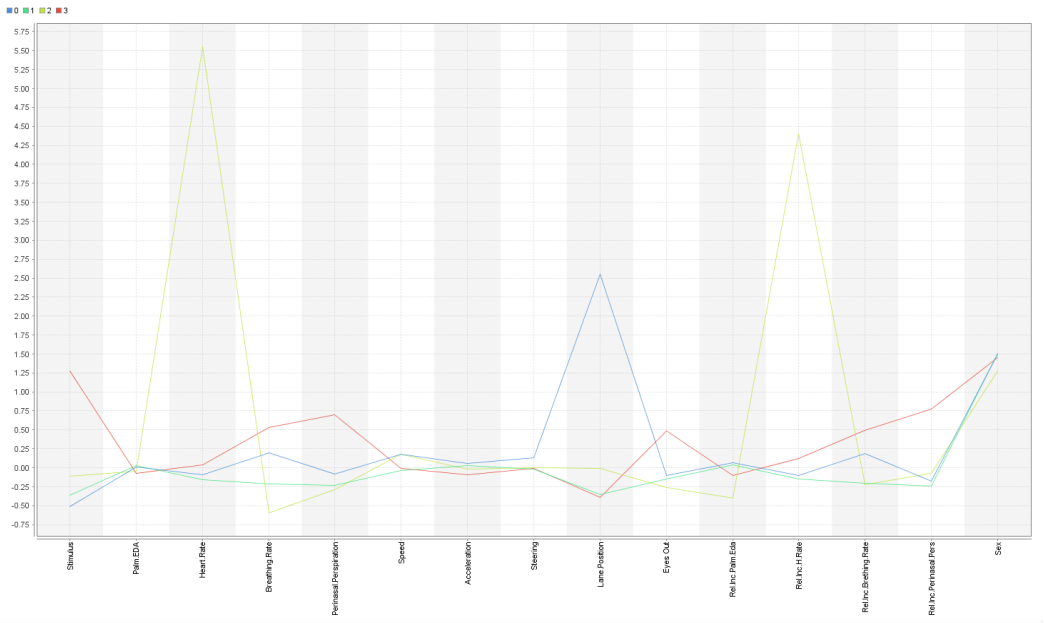


Fig. 40 Centroid chart of k-means, k=4 (own Elaboration, 2018)

As we can see in this analysis we have the same groups than for k=3, but with the adition of a new group that would be formed mostly by people who are not being distracted, the number of times they lose their sight of the road is lower than the average, and the variable Lane.Position tells us that in such instants the car was located more to the left of the limit of the lane than the other groups.

As it can be seen for K-means with two groups, the groups obtained are easily explained. However, as we increase the value of k, the groups obtained focus more on particularities of a single variable and less on the general behavior of the group. Validation techniques for clusterings will now be applied to determine the correct number of clusters based on statistical measures.

We will use 2 diferent index of internal validation for clustering algorithms. The first one will be the average distance within the cluster index, that is a measure of the cohesion of the groups [29], and the Davies-Bouldin index that measures at the same time, the cohesion between elements of the same cluster, and the separation between clusters [23].

K	average distance within the clusters	average distance within the cluster 0	average distance within the cluster 1	average distance within the cluster 2	average distance within the cluster 3	Davies-Bouldin Index
2	12,854	10,424	19,853	-	-	2,771
3	11,931	16,167	10,389	18,805	-	2,083
4	11,038	10,319	9,359	18,684	15,434	1,949

Table 9 clustering algorithms validation indexes (own elaboration, 2018)

As we can see, if we increase the value of k, the average distance within the clusters decreases, that means that more there is more cohesion between the elements of a given group and better the algorithm is. In the same way the Davies-Bouldin index decreases as we increase the value of k. That means that the objects within a group are more similar among them, and more different between groups. But that is only indication of the quality of the clustering algorithm and not the information retrieval capacity.

In the case of our clusters, it is determined that the most significant grouping is the K-means with k = 2, in which there is a group that receives distractions (its biometric measures increase due to the stress produced by attending several tasks at once and its attention to the road decreases) and a second group that does not receive distractions (their biometric measures do not increase because they do not respond to any external stimulus and their attention on the road is greater). As most of the distractions are oral or are done by the mobile phone, they cause a measurable biological response in drivers as a result of stress, and a loss in observable attention in the number of times they take their eyes off the road. But the responses are too similar to get groupings based on the distraction applied. To find out what distractions are more dangerous we will use the psicometric data of the volunteers.

Psicometric data analyses

Summarization

Row No.	Driving Test	average(Physical Demand)	average(Mental Demand)	average(Temporal Demand)	average(Performance)	average(Effort)	average(Frustration)
1	CD	7.970	14.701	10.239	9.918	14.246	10.082
2	ED	6.209	9.351	7.104	6.313	8.478	5.493
3	FDL	12.844	16.677	14.062	13.922	16.344	14.500
4	FDN	7.286	8.314	7.857	9.029	8.886	8.971
5	MD	11.478	14.224	11.306	11.381	14.836	11.567
6	ND	5.216	6	4.582	4.940	6.507	4.507
7	RD	5.269	7.642	5.075	5.388	8.246	4.955

Table 141 Average of the subjective opinions for each driving test (own elaboration, 2018)

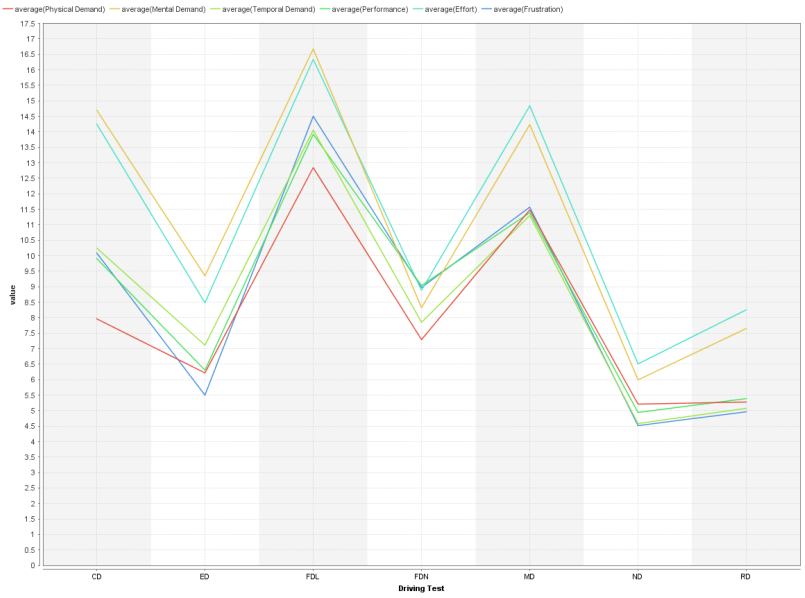


Fig. 44 Graph of the average of the subjective opinions for each driving test (own elaboration, 2018)

As it can be seen according to the graph shown, there are 3 tests that are scored with higher values (and therefore perceived as more difficult and frustrating):

- In first place the FDL test (Failed Drive Loaded) is classified with the biggest scores, much different of the FDN test (Failed Drive Normal), this must be caused by the saturation of information of the volunteers when they must focus on the screen.
- In second place the MD test (Sensorimotor Drive) where the drivers are distracted by phone calls or text messages. It receives the highest score for the single stimulus driving test. This must be related to the fact they are forced to lose sight of the screen while they use the mobile phone.
- In third place the CD test (Cognitive Drive), it obtains the higher scores in mental demand, and general high scores in the other atributes, but frustration values aren't as high as the other difficult test, this may be because the volunteers take it like a game.
- In the other hand we have the driving tests with low scores, where ED (emotional drive) have scores only slightly superior to the Normal drive test, and the relaxing drive.

K-means

Last we will proceed to make clusters with the subjective data from the NASA-TLX.

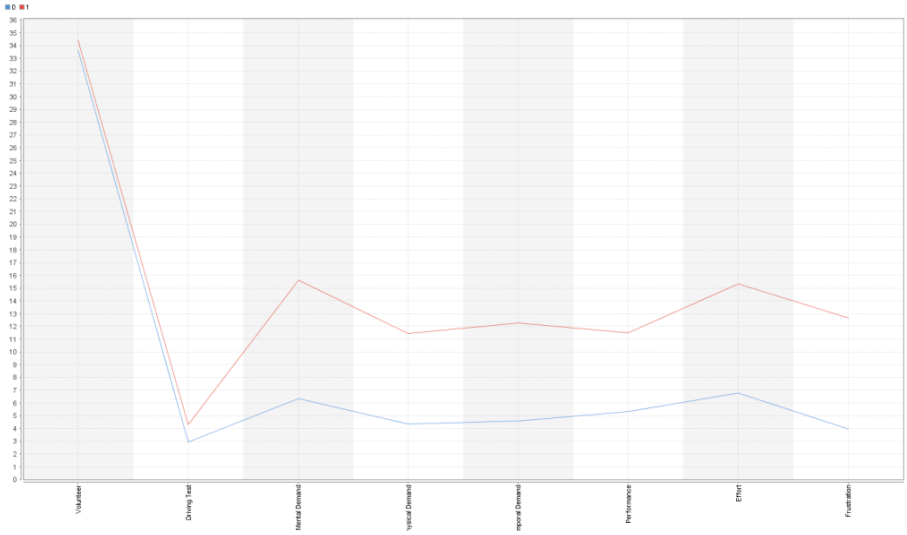


Fig. 45 Centroid chart for k-means k=2 using the psychometric data (own elaboration, 2018)

The algorithm classified the tests in two types, simple tests (mostly formed by the RD, ND, ED, and FDN tests) and hard tests (mostly formed by CD, MD and FDL) according to the score of the subjects. It can also be observed that when a test is considered difficult by a subject, all the attributes of the NASA-TLX scale are high even though they have no direct relationship between them. It is also notable that when a subject perceives the test as difficult and therefore the necessary effort is high, the performance that he perceives of himself is also superior. In the same way they also perceive a low performance of themselves when the test requires little effort.